University of Canterbury

Master Thesis

The effect of culling on breeding values

Author:

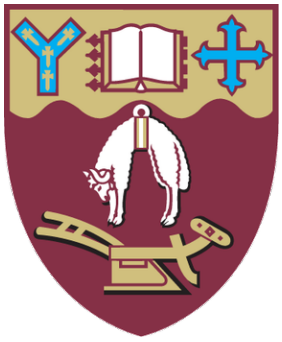
Yijia Qian

Supervisors:

Dr Daniel Gerhard

Dr Peter Jaksons

Dr Luis Apiolaza



A thesis submitted in partial fulfillment of the requirements

for the degree of Master of Science in Statistics

in the

School of Mathematics and Statistics

College of Engineering

2020

Abstract

Culling is inevitable in breeding trials. It removes individuals with undesirable traits but also introduces bias in breeding value estimation since we cannot ensure the population normality assumption. Here a stochastic simulation was used to study the impact of culling on the precision of breeding value estimation. Bivariate data were simulated for different combinations of genetic parameters such as a range of heritabilities for the traits and for a range of genetic correlation between the traits. This study has three goals: Quantify the effect of culling on the ranking order of genotypes based on their breeding values. Evaluate the potential of a using a multi-trait linear mixed model to reduce that bias introduced by culling. Last, test the efficiency of three potential trial management strategies which each retain a set percentages of culled individuals to minimise the bias in breeding value estimation.

The results showed that in the case one selects a small number of individuals (less than 10%) for a trait with a high heritability the impact of culling on the bias of BLUPs is much lower compared with trait with a lower heritability or when a larger number of plant is being selected for progression. The percentage of individuals culled did not influence the ranking persistence much compared with the other variables in the simulation study. Applying a multi-trait linear mixed model for breeding value estimation can reduce bias in breeding values due to culling. This approach shows a larger improvement when the correlation between the traits is high. All three trial management strategies appear to improve selection in terms of ranking persistence. However, retaining individuals systematically proves to be the most efficient trial management strategy.

This study provides practical information and advice for breeding programs, which deal with culling during the plant breeding process.

Acknowledgements

Over the past year, working on my thesis has provided me a more in depth knowledge of statistics and its real world applications especially in the filed of breeding. Working on simulation processes has improved my coding skills in R and opened up learning of new technologies such as GitHub and Latex. I have discovered so many R packages that I have never come across before. Working with the IT team has given me a peak into the technical aspects of managing such a project. I challenged and pushed beyond my comfort zone to present my findings in front of peers at university, breeders at Plant and Food Research institute and other academics. I discovered wearing an old pair of glasses during presentations helped me blur out the audiences and pretend to making eye contact. As a non-native English speaker, I encountered the many quirks of the English language. The phrase "short ans sweet" was my motto while constructing sentences. I realised short and simple sentences enabled me to convey my findings better than long complicated ones.

I'm grateful for the support I have received, without which this thesis would not have been possible. My thanks go to:

- My supervisors, Dr. Daniel Gerhard, Dr. Peter Jaksons and Dr. Luis Apoloiza. Thank you for all the encouragement, patience and academic advice given to me during my study and devoting the time to proofread my work. They provided me with guidance and confidence to attend various conferences and seminars to engage with delegates and present my work.
- Plant and food research institute. Thank you for providing me with the funding, platform and resources to achieve this study and giving me the opportunity to apply my work towards solving a real life issue.
- The IT team at Plant and Food Research. Special thanks to Irwin for providing all the necessary hardware so I was able to finish this thesis at home despite the lock-down.
- The department of Math and Statistics at the University of Canterbury. Thank you for the all the advice and banter from the staff and other postgraduate students while I was studying at the department.
- My family and friends. Special thank you to my partner Chris Li, for enduring my early morning alarms and frustrations during the past year.

Contents

List of Abbreviations

ANOVA	**AN**alysis **O**f **VA**riance
BLUE	**B**est **L**inear **U**nbiased **E**stimator
BLUP	**B**est **L**inear **U**nbiased **P**reditor
LMM	**L**inear **M**ixed **M**odel
MLE	**M**aximum **L**ikelihood **E**stimator
MR	**M**isclassification **R**ate
REML	**RE**stricted **M**aximum **L**ikelihood

Chapter 1

Background and Literature Review

The purpose of this thesis is to quantify the bias of breeding value estimation caused by culling in modern breeding trials. We also explore the possible improvement in breeding value estimation by incorporating the correlation between traits in multivariate linear mixed models (LMMs). Finally, we report on any potential improvements that could be made in breeding value estimation by modifying the trial management strategies.

This thesis is arranged as follow: In Chapter 1, we introduce some concepts of plant breeding and its history, the reasons of using breeding values in modern plant breeding, review the current literature regarding the relationship between selection and breeding value estimation bias, and review of a possible solution to reduce this bias. Chapter 2 summarises how breeding values are estimated and how the model is used for estimation. In Chapter 3 the data simulation process for this study is outlined. In the remaining chapters of this thesis, we show the summary of results (Chapter 4) and discuss our findings and possible future work (Chapter 5).

1.1 Selection in breeding history

Food and water are the most basic of human needs. Getting a stable supply has played an essential role in the establishment of human civilisation. The advent of cultivating crops has provided us with a steady supply of food. The lifestyle of our ancestors

changed from that of hunter-gatherers to that of agriculture and animal husbandry (Veatch-Blohm, 2007). It is believed that plant breeding has been part of agriculture for the last 10,000 years (Hallauer, 2011). It is the process of manipulating plant attributes, structure and composition to enhance the value of crops (Veatch-Blohm, 2007). During the development of agriculture, humans also discovered variety selection as a core concept of successful breeding (Veatch-Blohm, 2007). Selection could happen both naturally and artificially. When the environment determines a plants survival without human interference it is called natural selection. Plants successfully survive when they have adapted to their environment. Artificial selection occurs through human intervention. The selection criteria is not limited to the environment but often plants with desirable traits are less likely to survive in nature. Through human interference environmental effects can be reduced significantly for selected plants.

Traits recorded in the field or lab that are the visible expression of an individual's characteristic are called phenotype. Phenotypic selection has the most extended history in plant breeding and is used even today. An obvious way of selecting desired traits is based on phenotypes. It has been used as an early plant breeding unit of selection (Hallauer, 2011). Phenotypic selection has enabled humans to transition wild plant species to those that can be cultivated. It has proven effective especially in the case of maize (Duvick, 2005). Phenotypes are influenced by a combination of an individual's genetic makeup and its environmental effects (Falconer and Mackay, 1996). However, it is less affected by its environment (Hallauer, 2011). There is a continued interest in how to minimise the environmental effects to increase the effectiveness in phenotypic selection (Hallauer, 2011).

Phenotypic selection depends heavily on environmental effects. If we can quantify genetics and environmental effects and separate them it would be more effective when selection is only based on one's genetic effect. This genetic effect is called the breeding value. Fisher (1919) outlines the main principle of quantitative genetics. The work of Ronald Fisher and Gregor Mendel is the theoretical basis for many animal and plant breeding research developments (Hallauer, 2011; Veatch-Blohm, 2007; Falconer and Mackay, 1996). Gregor Mendel discovered the hereditary factors in plant traits and developed a theory that traits can be carried from parents to their offsprings through

genes. It was an important discovery as we could select and pass on desirable traits while avoiding to select those that are undesirable. This process could be streamlined over successive generations. With the quantification of genetics and development of breeding research, we can calculate the breeding values of a plant's specific trait. More importantly, we can cross two plants with high breeding values to create offsprings with an even higher breeding value.

In both natural and artificial selection, we indirectly used genotypes as the basic unit of selection. Even today with the development of molecular genetics and genetically modified organisms the genotype of an individual remains the unit of selection (Hallauer, 2011). The heritability of a trait is a key factor in determining how effective genotypic selection is. Heritability (h^2) indicates the percentages of total phenotypic variation due to genetic variation among individuals (Hill, 2010). If h^2 of a certain plant's trait is 0.5, it indicates that 50% of variation we observed is due to genetic differences between plants. Genomic selection is a more recent concept. It was first introduced by Meuwissen et al. (2001); their study concluded that the selection on genetic values predicted from markers can improve the rate of genetic gain in both animals and plants substantially. Various studies have proven the increased efficiency of breeding programs by means of genomic selection (Harris and Johnson, 2010; Song et al., 2019; Ceballos et al., 2016).

1.2 Linear Mixed model and breeding values

Before we get into the details of estimating breeding values, a brief introduction to the concept of genetic evaluation will help us understand how linear mixed models are constructed and why they are used in breeding value estimation. Genetic makeup can be divided into additive genetic effects and Mendelian sampling effects. Additive genetic effects are the parts of the phenotype one inherits from its parents, so one's additive genetic effect can be calculated from its parents. Mendelian sampling effects, however, are not predictable. They describe the variation between individuals caused by random allele combinations, which are broken in the gamete production and re-established in offsprings. Environmental effects not only depend on the location of where an individual has grown but also anything that could affect the performance

of an individual during its lifetime. Individuals with the same genetic makeup may perform quite differently due to the differences in their environment. It is assumed that the genetic makeup and environmental effects are independent from each other. Changes in the environment will not change one's genetic makeup and vice versa.

To understand the differences in variation within and between individuals, the separation of the underlying genetic and environmental components of phenotypic variation is crucial (Postma, 2006). In contrast with data collected in a laboratory, which is usually balanced, data collected in the field can be highly unbalanced. Furthermore, the genotypes in a breeding trial often come with a known but complex pedigree structure or with highly unbalanced data. Traditional methods such as ANOVA do not perform well with data analysis which incorporates such structure (Lynch et al., 1998). For the past few decades, LMMs have gained popularity in quantitative genetic research. They can make use all the information available in the pedigree, in particular the correlation structure between individuals. A linear mixed-effects model can also include a selection effect. This occurs both in nature and in artificial environments. Most importantly, it could include one or more environmental effects in the model (Kruuk, 2004). Researchers are trying hard to separate variation between genetic and environmental effects. The genetic effect of a given trait is inherited from its parents and can also be passed on to its offsprings. These effects are referred to as additive genetic effects or more commonly as breeding values. This separation makes the calculation of heritability (h^2) possible. LMMs not only make this separation possible, but they also make the separation possible of the two effects on both individual and population level. It quantified an individual's breeding value for a given trait which makes LMM an ideal tool for quantitative genetic parameters estimation (Postma, 2006).

There are several methods available to estimate the variance: i.e. analysis of variance (ANOVA), maximum likelihood (ML), restricted maximum likelihood (REML) and Gibbs sampling. The traditional ANOVA method uses least-squares to estimate variance for balanced data. Data from breeding trials can be unbalanced over time and space. Henderson (1953) outlined three ways to apply the ANOVA method to estimate variance from unbalanced data. ANOVA is easy to calculate and understand but this method also has its weaknesses. First, the variance estimates can become

negative. Second, the ANOVA estimator has an unknown distribution even under normality assumption. Third, ANOVA cannot be used for analytic comparison of different applications. These weaknesses makes ML the preferred method for unbalanced data (Searle, 1995). ML estimates the probability distribution parameter by maximizing the likelihood function. It demands that maximization be over the parameter space and overcomes the issue of ANOVA producing negative variance estimates for positive parameters. The asymptotic nature and assumption of normality of error terms makes ML an important estimation technique (Searle, 1995). The drawbacks of using ML is that when estimating the variance for a normal distribution the variance is a function of the actual mean. The actual mean is in fact unknown. A workaround is to replace the actual mean by the sample mean when estimating from a sample then dividing it by $N - 1$, which accounts for the uncertainty in the value of the actual mean (Weller, 2016). Patterson and Thompson (1971) proposed a way to analyze data collected from unbalanced block design. Today this method is known as Restricted or Residual maximum likelihood (REML). REML overcomes this issue by removing the mean (fixed effect) through a linear transformation (Weller, 2016). It is now widely applied in LMM variance estimation (Gurka, 2006; Corbeil and Searle, 1976; Liao and Lipsitz, 2002). Another way to estimate the variance in LMMs is by using a Bayesian framework. Given the development of computation power over the past few decades, Bayesian inference has increasingly played an important role in statistics. During the 1990s, Bayesian Markov Chain Monte Carlo methods were introduced in quantitative genetics (Sorensen et al., 1994). Bayesian Gibbs sampling technique usually apply when estimating breeding values as it could adapt to a high-degree of inbreeding and genotype-by-environmental interaction (Bauer et al., 2009). It has been practised heavily in animal breeding for multi-trait evaluation (Van Tassell and Van Vleck, 1996; Faria et al., 2007; Stock et al., 2008). Because we are simulating a general breeding and culling process, and no informative prior information is available, the difference between the results of REML and the Bayesian framework is assumed to be small. So in this study, we used a statistical package that fits LMMs using REML in the R environment called **ASReml-R**.

1.3 Culling in breeding trials

In breeding trials, breeders have different traits of interest at each stage. The process of culling, which is the removal of individuals i.e. plants or trees, happens over different stages according to different traits of interest, such as seeds failing to germinate, plants failing to pollinate, or the production of inadequate quality of offspring are removed at different breeding stages. It is a time and space saving strategy for breeders. If plants are removed in the early stages, the additional space could be used for breeding new varieties at a later stages. This saves the breeders' time in the field caring for plants and animals which will not produce offspring that satisfy our needs. Since culling occurs at each stage of the breeding trials, the data collected at the end of the trials contain genotypes with no information at all (failed to germinate), genotypes with incomplete information (culled before the end of trial) and genotypes with full information (survived the entire trial). The incomplete information for genotypes can be viewed as missing values. The assumption of normality no longer holds for data with missing values when estimating the breeding values and the selection of genotypes. One solution is to impute missing data using the related genotypes with complete information.

Gianola et al. (1989) classified missing data from animal breeding trials into three types. Although animal and plant breeding are different in many ways, the missing animal data classification shows some similarity to plant breeding. The study considered two correlated traits, where selection is based on the first trait. Type I is defined with all individuals having data available for the first trait, observations of the second trait are only available for selected animals. Type II: for selected animals, we have no observation for the first trait but have all for the second. The third type denotes individuals that have no observations recorded. An example of this can be seen in Swedish Landrace and Yorkshire breeding, where up to two-third of pigs were culled before any testing (Appel et al., 1998). For the first and second type of missing data, a single trait is used to decide whether an individual should be culled. Single trait analysis based on other traits will be biased as some of the information used in the selection process may not be available. This is often called culling bias (Mrode, 2014).

Henderson and Quaas (1976) proposes using a multiple-trait model to reduce the culling bias. Compared with a single-trait model, the multiple-trait model accounts for the correlation between traits and increases the estimation accuracy. Effects such as genotype-by-environmental effects and residual effects can be incorporated by the covariance structure in multiple-trait models (Pollak et al., 1984). Given different heritabilities of traits, low-heritability traits can achieve higher prediction accuracy by borrowing information from high-heritability traits by using a multiple-trait model in a simulated environment (Jia and Jannink, 2012). Similar results have been observed in research which dealt with real data (Volpato et al., 2019). The third type of missing data is more difficult to handle. Appel et al. (1998) showed multiple-trait analysis does not offer solution when all observations are missing. Unbiased predictions will be obtained if observations of all animals are available or if culling was done randomly. Both of these are hard to achieve in reality. Improving the analysis model is one approach to reduce culling bias while another approach is to adjust the data. In the horse racing industry, a large number of horses are culled before they enter a race. Klemetsdal (1992) suggested replacing the missing data by the average phenotypic observations of all the culled animals. A similar approach was suggested for swine breeding (Appel et al., 1998). Appel et al. (1999) investigated whether different culling strategies combined with data augmentation could decrease culling bias in swine breeding. Similar to racehorse breeding, pigs culled before testing have no phenotypic observations. This culling strategy was used to identify the low value animals between and within litters of pigs for swine breeding. Mehrabani-Yeganeh et al. (1999) demonstrates that in a two-stage selection program culling based on breeding values produces a more precise response than culling on phenotypes.

In this thesis, a simulation study using a single-trait analysis is performed to quantify the effect of culling on breeding values. A simulation study using a multiple-trait analysis is used to identify the improvements in breeding value estimation. We test whether trial management strategies could help reduce bias introduced by culling.

Chapter 2

Breeding Value Estimation

In this chapter we introduce the structure of LMMs (subsection 2.1.1) and how the variance of additive genetic effects and environmental effects is separated (subsection 2.1.2). The method used to estimate the variance of additive genetic effects and environmental effects is specified in section 2.2. Last, we introduce the structure of multi-trait LMMs (section 2.3).

2.1 Linear mixed models

Overview

A linear mixed model (LMM) is a parametric linear model that contains fixed effect parameters and random effect parameters. It is often used for the analysis of clustered data, repeated measurements and longitudinal studies. It was first introduced by Airy in 1956 (West et al., 2007) as an LMM with one random factor and no fixed factors. It has been developed over the years by various statisticians and is now widely used in medicine, social sciences and biology.

LMMs decompose the known effects into population effects and individual effects. Population effects are usually referred to as fixed effects and describe the relationship between dependent and independent variables for the entire population. In breeding analysis, they usually represent the average trait value of an individual's character among a population or a unit of analysis. Individual effects are usually referred to as random effects and describe the clusters or subjects within the population. Random effects

have many similar properties as the residual component in a regular linear model. It is the part which cannot be explained by the independent variable (fixed effect). LMMs allow us to look deeper into the residual of the model and separate it into the variances within and between individuals. Random effects can be separated into between individuals effects and within-individual effects. In breeding analysis, these are respectively referred to as the additive genetic effects or breeding values and non-additive genetic effect or environmental effect respectively. Breeding value refers to the value of an animal/plant in a breeding trial for a particular trait, it quantifies the individual's true genetic potential. Given the same environment, genotypes with higher breeding values should outperform those with lower breeding values; breeders would like to select plants carrying genotypes with higher breeding values through breeding trials.

$$y = X\beta + Zu + e \tag{2.1}$$

- y is a $n \times 1$ vector, n is the number of observations in data set.
- β is fixed effects matrix, it is a $p \times 1$ vector, p is number of levels for fixed effect.
- X is a $n \times p$ design matrix, for the fixed effects.
- u is a $q \times 1$ vector of breeding values, q is number of levels for breeding values.
- Z is a $n \times q$ design matrix for the breeding values.
- e is a $n \times 1$ vector of environmental effects.
- It is assumed that $E(Y) = X\beta$, $E(u) = E(e) = 0$.

Variance Components

As mentioned in subsection 2.1.1, random effects have similar properties to those of residuals in a regular linear model. Like residuals in linear mixed models it follows a normal distribution with mean 0 and variance V. LMMs partition the residuals into two components u and e, both following a normal distribution having mean 0 and variance G and R respectively. G equals to the numerator relationship A times the variance between breeding values (σ_a^2) and R equal to the identity matrix times the

variance of environmental effects (σ_r^2). Breeding values and environmental effect are assumed to be independent from each other, therefore, the covariance between the two effects is 0.

- $u \sim N(0, G),\ G = var(u) = A\sigma_a^2$, A is the numerator relationship matrix.
- $e \sim N(0, R),\ R = var(e) = I\sigma_r^2$, I is a $p \times p$ identity matrix.
- $$\begin{aligned} var(y) = V = V(\theta) &= var(Zu + e) \\ &= Zvar(u)Z' + R + cov(Zu, e) + cov(e, Zu) \\ &= ZGZ' + R \end{aligned}$$

The numerator relationship matrix A is added in G because the data collected from the breeding trials usually have subjects with missing phenotypic observations. This includes A in the variance among breeding values that allows us to borrow information from the relatives to fill these gaps. The individual's DNA is equally inherited from the male and female parent. One allele of each parent is randomly sampled from the two parents to create a new offspring. The individuals with mutual parents will have correlated breeding values. This relationship among individuals can be described by the numerator relationship matrix (A). It is a symmetric matrix where the diagonal elements of A are equal to one plus half of the relationship between its parents. The off-diagonal elements represent the relationship between two individuals which equals to half of the relationship between the individual and the parents of the other individual (Equation 2.2).

$$\begin{aligned} a_{diag} &= 1 + 0.5(a_{father,\ mother}) \\ a_{ind_1,\ ind_2} &= 0.5(a_{ind_1,\ father\ of\ ind_2} + a_{ind_1,\ mother\ of\ ind_2}) \end{aligned} \tag{2.2}$$

2.2 Parameter Estimation

In linear regression, parameters can be estimated using the standard least square method which minimizes the sum of squares of the residuals. However, this method

should not be applied to LMMs. The sum of squares of the residuals cannot be minimised because of it dependency on the variance-covariance structure of the residuals. This could cause estimation to fall out of parameter space and become negative. Another common estimation method is maximum likelihood estimation (MLE). Based on the density functions, MLE maximize the likelihood of parameter using an unbiased estimator, which gives a biased result with smaller variance. REML is an adaptation of MLE. It only maximizing the part of the likelihood which is location invariant (Searle, 1995). One of the reason why some statisticians prefer REML to MLE is because they transform the likelihood equation into a mean-free equation. REML accounts for the loss of degrees of freedom associated with the fixed effects in the model. LMM combined with REML is often use in breeding value estimation because data from breeding trials usually have missing values as a result of culling.

In LMMs, the dimensions and coefficients of the design matrix X and Z are specified by the designs β, u, e, G and R which are generally unknown. The estimators we use to estimate β and u are referred to as the best linear unbiased estimator (BLUE) and the best linear unbiased predictor (BLUP). These two methods are "best" in the sense that they have minimum mean squared error within the class of linear unbiased estimators and are unbiased in the sense that the average value of estimation equals to the average value of the quantity being estimated (E[BLUE(β)]=β and E[BLUP(u)]=u) (Robinson et al., 1991). Estimator and predictor are used to distinguish between fixed effect (estimator) and random effect (predictor).

The BLUE ($\hat{\beta}$) is

$$\hat{\beta} = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} y \tag{2.3}$$

Henderson (1963) showed BLUP ($\hat{u}$) as follow

$$\hat{u} = \hat{G} Z^T \hat{V}^{-1} (y - X\hat{\beta}) \tag{2.4}$$

Obtaining V^{-1} can be time consuming as y could have thousands of observations. Henderson (1963) suggested a method which allows us to jointly obtain $\hat{\beta}$ and $\hat{u}$ as

Equation 2.5

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ X^T R^{-1} Z & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix} \tag{2.5}$$

In this study, we choose REML to estimate the variance components. It transforms the likelihood equation into two statistically independent parts, Sy and Qy, where S is a projection matrix with rank $v = rank(X)$ and Q is a weighted matrix with rank v (Equation 2.6).

$$\begin{aligned} S &= I - X(X^T X)^{-1} X \\ Q &= X^T V^{-1} \end{aligned} \tag{2.6}$$

$E(Sy) = 0$ and such transformation of Sy is also referred to as error contrast. If $AA^T = S$, then

$$\omega = A^T y = A^T(X\beta + Zu + e) = A^T(Zu + e) \sim N(0, A^T V A) \tag{2.7}$$

Estimation of θ can now be made without including fixed effects which makes the estimation unbiased.

2.3 Multiple traits analysis

Characteristics of an individual and those between related individuals are often correlated. Measurement of one trait can give us information about the other correlated traits and accounting for this correlation can result in a more robust and precise analysis of the available data. Multivariate information can be incorporated into a LMM. We can simply consider multi-trait models as a stack of univariate models for n traits. The model for each trait has been given in Equation 2.1. Consider a multi-trait model for K traits where the model for the kth trait can be written as

$$y_k = X_k b_k + Z_k u_k + e_k \tag{2.8}$$

Where k is an index to indicate the trait ($k = 1,\ 2,\ \ldots,\ K$)

Model for the multivariate analysis model can be written as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_K \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix} + \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ K0 & 0 & \cdots & Z \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_K \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_K \end{bmatrix} \tag{2.9}$$

- y_K is a vector of observation for the ith trait.
- X_K is design matrix for the fixed effect.
- b_K is a vector of fixed effect for the ith trait.
- Z_K is design matrix for additive the genetic effect
- u_K is a vector of additive genetic effects for the ith trait.
- e_K is a vector of environmental effects for the ith trait.

The environmental effect e, follows a multivariate normal distribution with mean 0 and variance R. R is a matrix, formed by the Kronecker product of the variance-covariance matrix E and the identity matrix I. E is a $m \times m$ matrix that represents within-individual environmental effects where K is the level of traits in the data. As shown in 2.11, the Kronecker product is used here to allocate the covariance to all individuals.

$$e \sim MVN(0, R) \;\; where \;\; R = E \otimes I \tag{2.10}$$

$$R = E \otimes I = \begin{bmatrix} \epsilon_1^2 & \epsilon_{12} & \epsilon_{13} & \cdots & \epsilon_{1k} \\ \vdots & \epsilon_2^2 & \epsilon_{23} & \cdots & \epsilon_{2k} \\ \vdots & \vdots & \epsilon_3^2 & \cdots & \epsilon_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \epsilon_k^2 \end{bmatrix} \otimes I = \begin{bmatrix} I\epsilon_1^2 & I\epsilon_{12} & I\epsilon_{13} & \cdots & I\epsilon_{1i} \\ \vdots & I\epsilon_2^2 & I\epsilon_{23} & \cdots & I\epsilon_{2i} \\ \vdots & \vdots & \epsilon_3^2 & \cdots & I\epsilon_{3i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \cdots & I\epsilon_{ii}^2 \end{bmatrix} \tag{2.11}$$

The additive genetic effect variance can be constructed under a similar argument. The random effects u_j follows a multivariate normal distribution with mean 0 and variance G. G is constructed by variance-covariance matrix C and relationship matrix A. C is a $m \times m$ between-individual additive genetic effects where m is the level of additive genetic effect. The diagonal elements in E are the variance of additive genetic effect between genotypes and off-diagonal elements are the correlation/covariance between traits on a genotype level. A is the numerator relationship matrix as discussed in subsection 2.1.2

$$u \sim MVN(0, G) \ \ where \ \ G = C \otimes A \tag{2.12}$$

$$G = C \otimes A = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1j} \\ \vdots & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2j} \\ \vdots & \vdots & \sigma_3^2 & \cdots & \sigma_{3j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \sigma_j^2 \end{bmatrix} \otimes A = \begin{bmatrix} A\sigma_1^2 & A\sigma_{12} & A\sigma_{13} & \cdots & A\sigma_{1j} \\ \vdots & A\sigma_2^2 & A\sigma_{23} & \cdots & A\sigma_{2j} \\ \vdots & \vdots & A\sigma_3^2 & \cdots & A\sigma_{3j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \cdots & A\sigma_j^2 \end{bmatrix} \tag{2.13}$$

The mixed-model equation can be written as

$$\begin{bmatrix} X^T(E^{-1} \otimes I)X & X^T(E^{-1} \otimes I)Z \\ Z^T(E^{-1} \otimes I)Z & Z^T(E^{-1} \otimes I)Z + G^{-1} \otimes A^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T(E^{-1} \otimes I)y \\ Z^T(E^{-1} \otimes I)y \end{bmatrix} \tag{2.14}$$

In this study, we consider data sets with two traits so that the genetic effect variance-covariance matrix C (Equation 2.15) and environmental effect covariance matrix E (Equation 2.16) are both a 2×2 matrix.

$$G = C \otimes A = \begin{bmatrix} \sigma_{g_1}^2 & \rho_g \sigma_{g_1} \sigma_{g_2} \\ \rho_g \sigma_{g_1} \sigma_{g_2} & \sigma_{g_2}^2 \end{bmatrix} \otimes A \tag{2.15}$$

$$R = E \otimes I = \begin{bmatrix} \sigma^2_{r_1} & \rho_r \sigma_{r_1} \sigma_{r_2} \\ \rho_r \sigma_{r_1} \sigma_{r_2} & \sigma^2_{r_2} \end{bmatrix} \otimes I \tag{2.16}$$

Chapter 3

Simulation Study

There are two key advantages to using simulation studies in this thesis. First, the real breeding values are unknown which makes it hard to measure how selection would affect breeding values. Second, breeding trials are expensive, both in terms of cost and time. Simulation studies overcome these issues while allowing us to evaluate the effects of selection under variable conditions. Stochastic simulation was used to generate phenotypic data for two correlated traits (T_1 and T_2). Mean of dry-matter (20.57) and soluble-solids (17.39) from a real data of a kiwifruit breeding trial was used as a reference here to simulate our data.

In this thesis we will investigate the following three research questions:

1. Does culling have an effect on estimating BLUPS?

2. Can we use the known or estimated correlations between traits in a multivariate analysis to reduce the observed bias in BLUPs introduced by culling?

3. Can we reduce the culling bias by retaining a portion of the culled plants?

The data generation process is explained in section 3.1; it includes the process of generating the pedigree (subsection 3.1.1), environmental effects, true breeding values and the phenotypic data per plant (subsection 3.1.2). The culling process is explained in section 3.2. The model fitting process is explained in section 3.4. The method of choice for presenting the results of the simulation study is explained in section 3.5. Last, section 3.6 specifies the parameter settings for each simulation study.

3.1 Data generation

Creating the pedigree

We used a three-generation pedigree structure for our simulation studies. A reduced pedigree example is shown in Figure 3.1. The pedigree includes three generations of plants in which only the phenotypic data from the third generation is used. The first generation has five crossings of plants and each group contains one male and one female plants with an unknown parent. Each group has five male offspring selected. Each of these male offspring were then randomly crossed with the same five female plants who have unknown parents. Ten offspring produced from each of these crossings were selected, five females and five males. In total, the population consists of 1290 genotypes over the three generations, each represented by two plants from a total of 2580 plants.

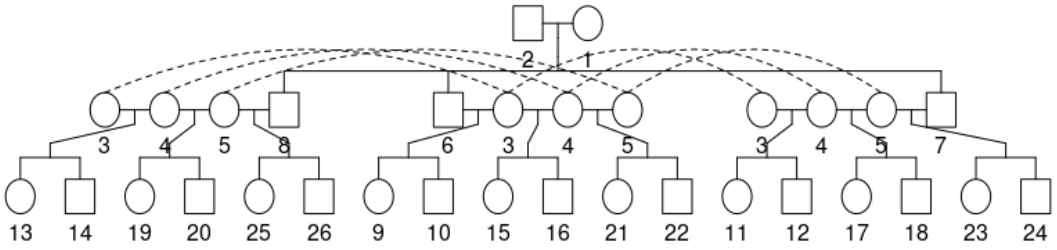


FIGURE 3.1. Schematic example of a three generation population structure. One crossing in the first generation, three male offspring selected and each randomly crossed with three females who have unknown parents, two offspring, one male and one female from each crossing.

Simulation of records

The Simulation process used in this section refers to the method used in Appel et al. (1995). The vector of environmental effects e sampled from a multivariate normal distribution with mean 0 and variance R (Equation 2.16), is given as $e_i = L'_e x$, where e_i is a vector of the environmental effect for plant i, L'_e is the lower triangular matrix, resulting from a Cholesky decomposition of the variance-covariance matrix $E = L'_e L_e$

and x is a matrix of random normal deviates sampled from a normal distribution with a mean of 0 and variance of 1.

Founders are those plants that have unknown parents. Neither pedigree information nor phenotypic data is available for them. True breeding values of founders were sampled from a bivariate normal distribution with a mean of zero and variance of G (Equation 2.15), as $g_i = L_c' x$, where g_i is a vector of the true breeding values for plant i, L_c' is the lower triangular matrix resulting from a Cholesky decomposition of the variance-covariance matrix $C = L_c' L_c$ and x is a vector of random normal deviates sampled from a normal distribution with mean of 0 and variance of 1. We are interested in finding out how culling affects traits with different levels of heritability. The variance of true breeding values for trait T_1 are therefore generated according to a heritability ranging from 0.1 to 0.9 as $h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$. The true breeding values of the founder's offspring were generated according to Equation 3.1. g_i is a vector, which contains the true breeding value for each offspring, g_{father} and g_{mother} are vectors of the true breeding values of the parents. m_i is a vector of the Mendelian sampling terms. For each genotype, the Mendelian sampling term is calculated as $m_i = \sqrt{0.5 \times C}$.

$$g_i = 0.5 g_{father} + 0.5 g_{mother} + m_i \tag{3.1}$$

Phenotypic observations for each plant were simulated as depicted in Equation 3.2, where y_i is a vector of phenotypic observations for both traits, a_i is a vector of the true breeding value of the plants genotype and e_i is a vector of environmental effects.

$$y_i = m_i + a_i + e_i \tag{3.2}$$

As mentioned in section 2.3, the plant effects G and R follow a multivariate normal distribution. There are pre-existing commands in statistical computing software (such as R) which allows us to sample directly from a multivariate normal distribution. The issue with these commands is that they are time-consuming. The data used in breeding analysis includes high dimensional relationship matrices. This makes the decomposition

process quite computational expensive and more importantly, the Cholesky decomposition has to be done repeatedly during the simulation. The method used in this study allows Z_r and Z_g sample in multiple iterations with just one Cholesky decomposition, which makes the process computationally much cheaper and thus faster. The Cholesky decomposition is crucial to this simulation and is challenging due to the dependency between plants, meaning one cannot be generated without considering the other. The Cholesky decomposition transforms the correlated variables into an uncorrelated space by mapping the variables with a covariance matrix centred at the origin.

3.2 Culling process

The observations with missing phenotypic data as a result of culling were generated by deleting records from the simulated data sets. In Figure 3.2, we demonstrate what could happen to the data when culling occurs. The X-axis is the phenotypic value of dry matter for a sample of fruit, the Y-axis is the phenotypic value of the soluble solids concentration of the sample of fruit. In this case the two traits are highly correlated. Assume culling would based on dry-matter, all plants which produced fruits with an average dry-matter value lower than 22 would be culled. While culling happens, traits that are correlated with dry-matter, such as soluble-solid will not be consider in the decision of the process, however the data for these traits are affected as a by product of culling.

In breeding trail, data collection and culling could happen at many stages according to different trait, for example, some phenotypic data are collected in the field, because are easy to measure (eg. weight), then fruits or crops are taken into the lab for further testing. Individuals failed to pass the threshold for lab test are then culled, which left us a data with a full set of phenotypic data for trait measured in the field and a incomplete set of phenotypic data for trait measured in the lab. Another situation is when culling happened in the filed, if plants fail to reach the threshold of culling for our trait of interest, then the entire plant will be removed (plant did not pollinate or have sign of contagious disease), no phenotypic information will be recorded for this plant. In this thesis and the following simulation studies we assume that plants are culled only based on trait T_1. Data sets were arranged in ascending order, according

to the phenotypic observation of T_1. Plants with the lowest values for the phenotypes of T_1 were culled first. We simulated the affect of the culling on the available data for the following two scenarios:

- Scenario One. The culled plant has some percentage of one phenotypic observation (T_1) missing and only has phenotypic observations of T_2.
- Scenario Two. A certain percentages of plants are removed completely. The removed plants are excluded, while remaining plants having both observations of T_1 and T_2.

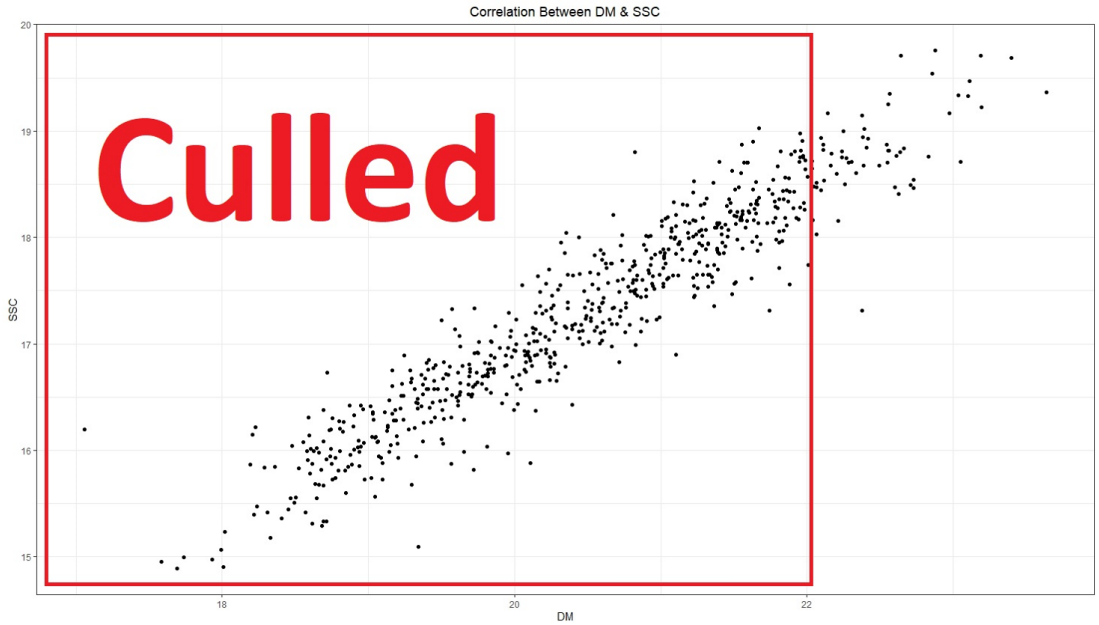


FIGURE 3.2. Example of how a culling process can affect the soluble solid concentration trait data when culling is based on another trait, in this case dry matter value. Note, in this case all data is missing as described in scenario 2

3.3 Trial management

The idea of trial management is to retain a percentage of plants from the culled population. The retained plants were sampled by three different methods.

- Random method. In this scenario we assume that the phenotypic observations of the plants are uniformly distributed. All the discarded plants have the same probability to be retained and are thus randomly selected.

- Sequential method. We assume that the phenotypic observations are normally distributed. The retained plants are sampled from one side (in our case the lower tail) of the distribution. This is done sequentially in such a way that plants with the smallest phenotypic value are retained first.

- Systematic Method. We assume that the phenotypic observations are normally distributed. The retained plants are sampled from the lower tail systematically and plants were sampled sequentially. However if two plants from the same genotype are retained during this process then the second plant with the higher phenotypic observation is no longer retained. Given the percentage of plants to be retained this method ensures we have a more diverse pool of genotypes at the end of the trial.

3.4 Model fitting & parameter estimation

Data sets used to quantify culling effect only contained phenotypic observations and true breeding values of T_1, they are fitted in an univariate LMM where the observations of T_1 are the response variable and the population mean as a fixed effect. Both variances of breeding values and residuals are assumed to follow a heterogeneous variance model. While data sets used for the multiple trait analysis and trial management strategy assessment include phenotypic observations and true breeding values of T_1 and T_2. Data set were fitted using a bivariate LMM, in these cases the observations of T_1 and T_2 are the response variables and the population means of each trait are the fixed effects (fixed intercepts). The variances of the random and fixed effects follow a correlation model where $C_{ii} = 1$, $C_{ij} = \phi_{ij}$, $i \neq j$, $|\phi_{ij}| < 1$. Both culling scenarios are based on the phenotypic observations of T_1 only. In our study, we assume plants that for all traits, higher value for the traits are preferred compared with lower values: e.g. a higher Soluble solid concentration is preferred over a lower concentration. For each scenario, we assume a certain percentage of data points were removed due to the culling process, we call this $p\%$. Plants with lower T_1 observations were either part of a group where a certain percentage of T_1 observations were removed (Scenario One) or where both T_1 and T_2 observations were removed. (Scenario Two), we refer this percentage as $s\%$.

In this study, breeding values and residuals were estimated using ASReml-R. The software package is designed to fit a general LMM to a moderately large data with complex variance models. The computation efficiency arises from using the Average Information REML algorithm and sparse matrix operations (Gilmour et al., 2009).

3.5 Presenting results

The results from the simulation studies are presented based on the consistency of genotype ranking before and after culling. Before culling, genotypes were arranged in descending order according to their true breeding values. We call this ranking the true ranking. Breeding values estimated from LMM were also arranged in descending order after culling, this ranking is referred to as the estimated ranking. The consistency of the ranking was examined for top n percentage of genotypes (top-n genotypes $n\%$) with the highest true ranking. In practise, the top-n genotypes are those genotypes that would progress to the next trial stage. We evaluate the proportion of genotypes that are in the top-n genotypes based on the true ranking but are no longer classified as such in the model estimates. This is called the misclassification rate (MR). For example, if genotype A was ranked 40^{th} by true ranking, but after culling, its ranking dropped to rank 65, we can no longer select it as only the top 50 are selected for the next trial stage, then genotype A has been misclassified. If genotype A was ranked number 49, then it would not have been misclassified since it was still included in the sample. The number of misclassified genotypes divided by the total number of genotypes in the top-n ranking is called the misclassification rate.

3.6 Simulation setting

In this study, a total of 1500 iterations were simulated for each question. In each iteration, the simulation starts with generating a variance-covariance matrix for the environmental effect by specifying variance and correlation between residuals of both traits. From the residual variance and heritability, the variance-covariance matrix for the additive genetic effects can be specified. Variance of breeding values for both traits were simulated according to heritability. For the first trait T_1 the heritability ranged

from 0.1 to 0.9, T_2 heritability had a fixed value, correlation between breeding values ranged from 0 to 0.9 except when quantify the the culling effect.

Phenotypic observations can be simulated given the variance-covariance matrix of the additive genetic effects and environmental effects. In each set of simulations of phenotypic data, we looked at a set of culling percentages ($c\%$). The MR was calculated for each culling percentage given a certain percentage of top-ranked genotypes, we refer this percentage as $s\%$.

In Question One, for each iteration, there were 9 data sets generated where 30 unique MRs were calculated given different combinations between culling percentage and selected number of top-ranked genotypes.

In Question Two, for each iteration, there were 36 data sets generated where 60 unique MRs were calculated given different combination between culling percentage and number of top-ranked genotypes for both univariate and multivariate model.

In Question Three, for each iteration, there were 36 data sets generated where 108 unique MRs were calculated given different combinations between culling percentage and number of top-ranked genotypes for four multivariate models. The settings for all parameters in each question are specified in Table 3.1 and Table 3.2.

Question	h	σ_{g_2}	γ_g	σ_{r_1}	σ_{r_2}	γ_r
1	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9	3	0	1	1	0
2	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9	3	0, 0.1, 0.5, 0.9	1	1	0.9
3	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9	3	0, 0.1, 0.5, 0.9	1	1	0.9

TABLE 3.1. Parameter setting of p, n, s and number of iterations

Question	p (%)	n(%)	s(%)	iterations
1	0, 20, 40, 60, 80	1, 5, 10, 20, 30, 40	NA	1500
2	0, 20, 40, 60, 80	1, 5, 10, 20, 30, 40	NA	1500
3	10, 30, 50	1, 5, 10, 20, 30, 40	10, 30, 50	1500

TABLE 3.2. Parameter setting of p, n, s and number of iterations

Chapter 4

Results

In this chapter, we summarise the result of culling effect on breeding values estimation accuracy by looking into the ranking consistency of genotypes before and after culling in section 4.1. The comparison between the accuracy of breeding value estimation by multi-trait LMM with and without specification of the correlation structure in heterogeneous variance model is summarised in section 4.2. The results for the improvement on breeding value estimation by the three trial management strategies are summarised in section 4.3.

4.1 Quantify culling effects on breeding values

Before looking at the summary result of 1500 iterations, we illustrate the culling effect on the true- and estimated genotype ranking from a single iteration. In order to observe changes of all genotypes, a reduced pedigree is used here to generate the true breeding values. There are 3 pairs of individuals with unknown parents in the first generation. Three male offspring from each parents are selected to cross with a female individual whose parents are unknown. From this cross, three female offspring and three males offspring are selected. There are 54 genotypes simulated in the third generation. In each iteration step, we arranged all simulated genotypes according to their breeding values in descending order and record their true ranking. BLUPs estimated by a univariate model from data with missing observations are arranged in descending order to obtain the estimated ranking. In this single iteration, traits with heritability of 0.1 and 0.9

and culling percentages of 10% and 80% are simulated; the results are summarised in Figure 4.1.

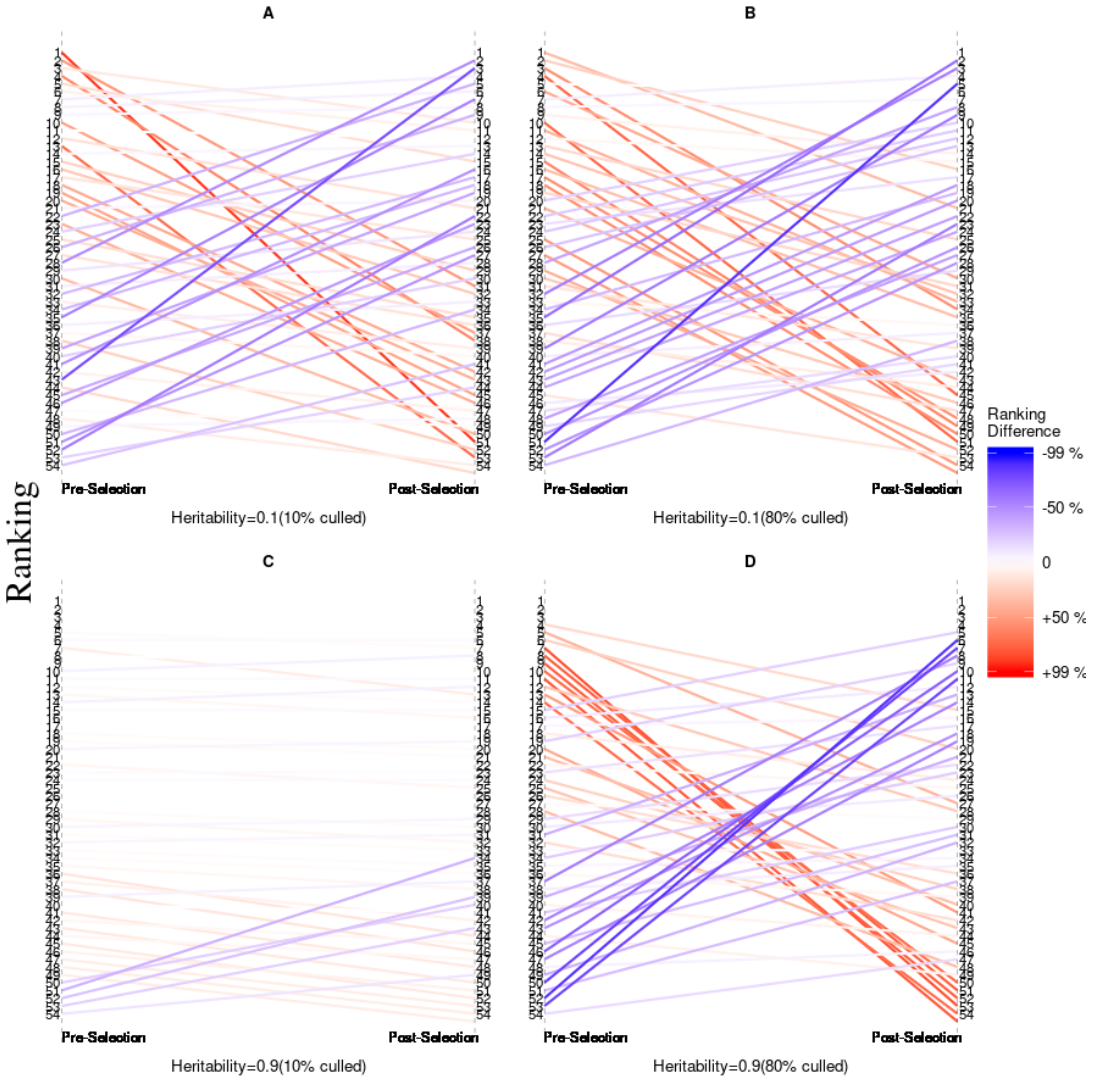


FIGURE 4.1. Ranking difference before and after culling. Plot A shows 54 genotypes ranking changes with heritability equals to 0.1 with 10% subjects removed. Plot B shows 54 genotypes ranking changes with heritability equals to 0.1 with 80% subjects removed. Plot C shows 54 genotypes ranking changes with heritability equals to 0.9 with 10% subjects removed. Plot D shows 54 genotypes ranking changes with heritability equals to 0.9 with 80% subjects removed.

In Figure 4.1, the numbers on the left y-axis indicate the true ranking of all 54 genotypes with in ascending order. The right y-axis shows the post-selection ranking of all genotypes in the same order. Lines connected the two numbers indicate that these two rankings belong to the same genotype. The gradient line colour suggests the percentage changed in the ranking. Line colour changes from blue to red as ranking difference

increase from −99% to 99%, brighter colour indicates large changes, lighter colour indicates otherwise. If the observed change is zero or near zero between the two rankings then the lines is white. For example, in plot Figure 4.1(A), genotype ranked number one at true ranking had a significant drop in the estimated ranking (number 51) after 10% subjects removed, so the line connects the two number is in bright red.

We can see that when the heritability is low (Figure 4.1 A & B), the changes in true- and estimated ranking are large regardless of the culling percentage. In this case removing either 10% or 80% subjects both cause dramatic ranking changes. When the heritability of a trait is high, the removal of 10% of the individuals has a very small impact on the genotype ranking, as shown in (Figure 4.1 (C)). Lines connect ranking numbers can barely be seen in the plot as most of the rankings stay the same or did not change much. The impact on the ranking starts to show when more individuals are culled. In Figure 4.1 (D), genotypes with low true ranking are now ranked at the top in estimated ranking, on the other hand, genotypes with high true ranking were ranked at the bottom at the estimated ranking. In reality, we would have had a loss in either situation, but the latter will trigger a much larger financial loss. If genotypes with low true ranking were selected (these are genotypes connected with bright blue colour), it is likely before the new variety made to the market, the genotype will be removed at a later trial stage due to poor performance. But if we miss out on genotypes with high true ranking (genotypes connected by red lines), we are missing out on genotypes, which could potentially be a commercial success.

Figure 4.2 shows the MR of the population of genotypes given different culling percentages and various levels of heritability. On the x-axis the heritability ranges from 0.1 to 0.9. On the Y-axis the MR ranges from 0% to 100%. Each panel shows the results for a specific proportion of genotypes retained at the end of the breeding trial. The line colour indicates different culling percentages i.e. 80%, 60%, 40%, 20% and 0% (or no culling present).

When retaining top 1%&5% of plants

- All MR curves follow a downward trend and are independent of the culling percentage.

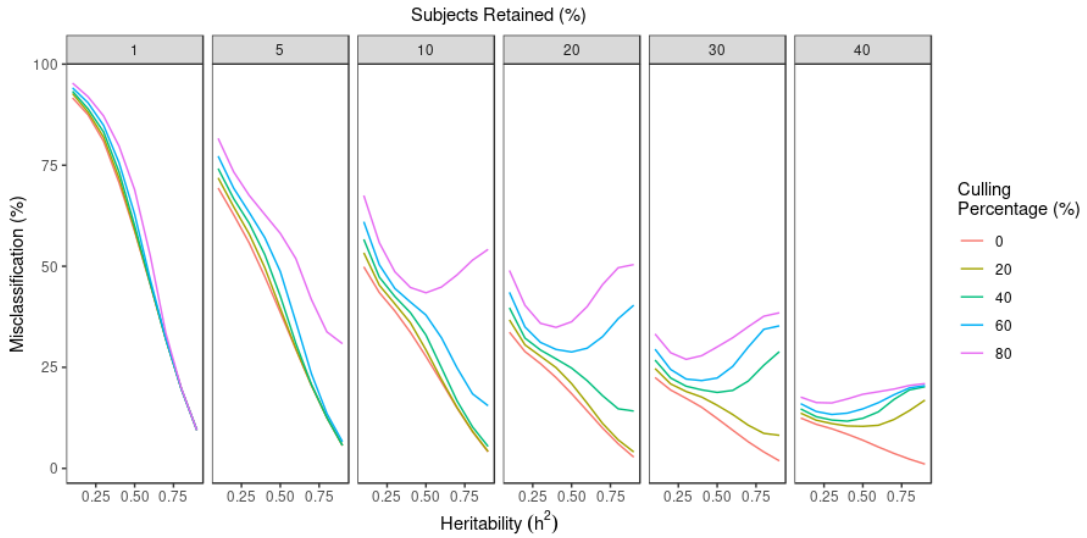


FIGURE 4.2. Misclassification rate calculated using a univariate model for T_1 given combination of different culling percentage, heritability and amount of plants retained.

- The variance of MR values is higher at lower heritability (0.1) but tapers to a small band as heritability increases with the exception of culling 80% of subjects when retaining top 5% plants .

When retaining top 10% of subjects

- The MR curve follows a concave pattern when 80% of the subjects are removed.
- When 80% of subjects were removed we observed the lowest MR when heritability is 0.5.
- The MR curves follow a downward trend for the other culling percentages.

When retaining top 20% of subjects

- The MR curve follows a concave pattern when 60% and 80% of the subjects are removed.
- When 80% of subjects were removed we observed the lowest MR when heritability is 0.4. When 60% of subjects were removed we observed lowest MR when heritability is 0.5.
- The MR curves follow a downward trend for the other culling percentages.

When retaining top 30% of subjects

- The MR curve follows a concave pattern when 40%, 60% and 80% of the subjects are removed.
- When 80% of subjects were removed we observed the lowest MR when heritability is 0.3. When 60% of subjects were removed we observed lowest MR when heritability is 0.45. When 40% of subjects were removed we observed lowest MR when heritability is 0.46.
- The MR curves follow a downward trend for the other culling percentages.

When retaining top 40% of subjects

- The MR curve follows a concave pattern of all the other culling percentage but 0%
- The lowest MR observed for 80%, 60%, 40% and 20% culling percentages are when heritability is 0.25, 0.3, 0.45 and 0.6 respectively.
- The curves observed at 60%, 40% and 20% are shallow compared to those that retain a lower proportion of top subjects.

4.2 Multi-trait approach for BLUP

In this section, we compare the breeding value estimation accuracy in terms of genotype ranking precision by adding a correlation relationship to traits at the level of additive genetic effects and environmental effects in a multi-trait LMM. The same data was fitted for two heterogeneous variance multi-trait models. One assumes there is no correlation (Zero-cor) while the other assumes a correlation between additive genetic effect and environmental effect (With-cor). Result of plants culled with T_1 observation (culling scenario one) is summarised in subsection 4.2.1. Results for both observations culled per plant are summarised in **??**. MR for breeding values, estimated from multi-trait LMMs that assume no correlation structure, were compared with those that have a correlation structure. This was done to depict the effect of assuming a correlation structure into the estimation process.

Plants with one observation missing

In a Zero-cor model, we assume zero correlation between traits on breeding value level. This feature is the reason we do not observed MR varying much given the different correlation between breeding values. In Figure 4.3 the heritability ranges from 0.1 to 0.9 and is measured on the x-axis. The top-n percentages of plants retained is indicated by the panel number above. The MR is measured on the y-axis. The true correlation between breeding values is indicated by the panel number on the right-hand side. Correlation between environmental effect is a fixed value (0.9). The line colours indicate different culling percentages and the line types indicate different models.

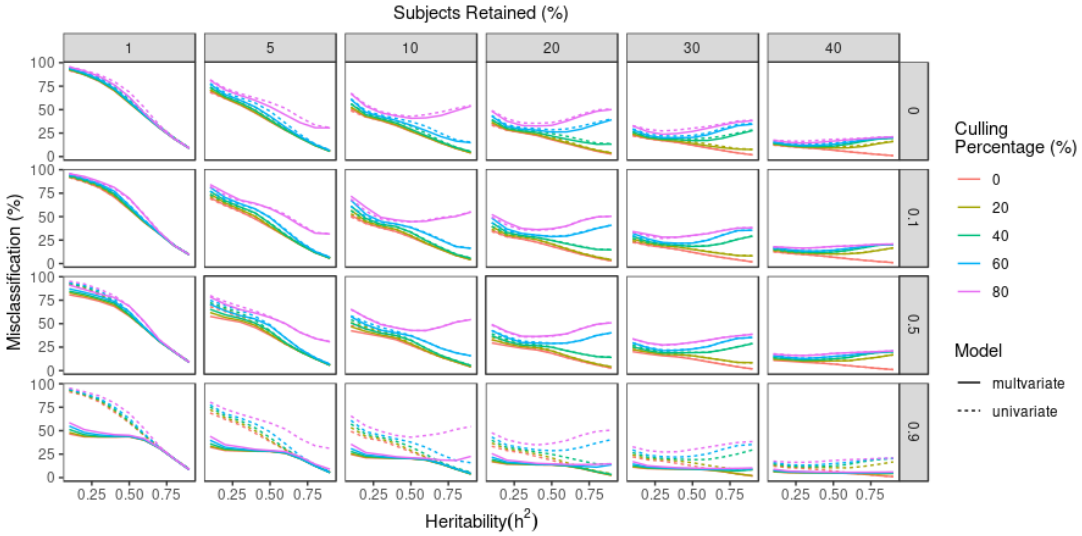


FIGURE 4.3. Misclassification rate calculated using multi-traits models when one phenotypic observations of subjects are missing.

When retaining the top 1% & 5% of plants:

- All MR curves follow a downward trend and are independent of the culling percentage for both models when correlation between breeding values is low (0.1) or moderate (0.5).
- MR curves follow a downward trend and are independent of the culling percentage for Zero-cor models when correlation between breeding values is high (0.9).
- In with-cor models, when the correlation between breeding value is high (0.9), the misclassification rate remains stable for lower to moderate heritability (from 0 to 0.5), but then proceeds to follow a downward trend.

- The variance of MR values is higher at lower heritability (0.1) but tapers to a smaller band as heritability increases for both models with the exception of Zero-cor model when 80% of plants are removed while retaining top 5% individuals.

- With-cor model shows a gradient improvement in MR compared to Zero-cor model as correlation increase when heritability is low (0.1) to moderate (0.6). This improvement is no longer present when heritability is higher than 0.6

When retaining top 10% of plants:

- The MR curve follows a concave pattern when 80% of the plants are removed for both models.

- In with-cor models, when correlation between breeding value is high (0.9), the MR remains stable for low to moderate heritability (from 0 to 0.6) but then proceeds to follow a downward trend with the exception of removing 80% of plants.

- When 80% of subjects were removed we observed the lowest MR when heritability is 0.5 with no correlation between breeding values. The lowest MR did not change much when correlation increased.

- The MR curves follow a downward trend for the other culling percentages.

When retaining top 20% of plants:

- The MR curve follows a concave pattern when 60% and 80% of the plants are removed for both models with the exception of With-cor models with high correlation (0.9).

- In with-cor models, when correlation is high (0.9), the MR remains stable as heritability increase for all culling percentages.

- In Zero-cor model, When 60% and 80% of subjects were removed the lowest MR we observed did not change much even when correlation varied.

- The MR curves follow a downward trend for the other culling percentages.

When retaining top 30% of plants:

- The MR curve follows a concave pattern when 40%, 60% and 80% of the plants are removed for both models with the exception of With-cor model that have a high correlation (0.9).

- In with-cor models, when correlation is high (0.9), the MR remains stable as heritability increase for all culling percentages.

- In Zero-cor model, When 40%, 60% and 80% of plants were removed the lowest MR we observed did not change much as correlation increased.

- The MR curves follow a downward trend for the other culling percentages.

When retaining top 40% of plants:

- For null (0) to moderate correlation (0.5), we observed MR follows a downward trend when no culling occurs otherwise has a very shallow concave pattern for both models.

- For high correlation (0.9), we observe a downward MR trend when culling percentage is low (0% to 20%) for both models while the others follow a very shallow concave pattern.

Remove full subject

A summary plot of the changes in breeding values MR estimated from Zero-cor LMMs and With-cor LMMs having different correlation between additive genetic effect and different heritability after completely removing an individual is presented in Figure 4.4. The Y-axis indicates the MR and the x-axis indicates heritability. Figures on the right hand side of the panel indicates correlation between the true additive genetic effect. The top-n percentages of plants retained is indicated by the panel number above. The correlation between environmental effects is fixed at a value of 0.9. The line colours indicate different culling percentages and the line types indicate different models.

When retaining top 1% & 5% of plants:

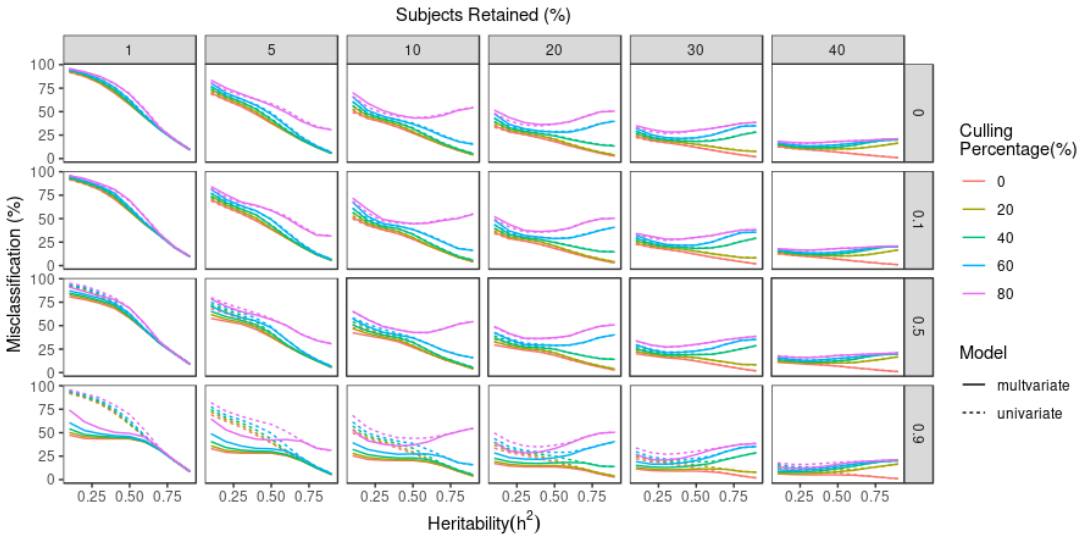


FIGURE 4.4. Misclassification rate calculated using multi-traits models when all phenotypic observations of subjects are missing.

- All MR curves follow a downward trend and are independent of the culling percentages for both models when correlation between breeding values is low (0.1) to moderate (0.5).

- MR curves follow a downward trend and are independent of the culling percentage for Zero-cor models when correlation between breeding values is high (0.9).

- In with-cor models, when correlation between breeding value is high (0.9), the misclassification rate remains stable for low to moderate heritability (from 0 to 0.5) but then proceeds to follow a downward trend.

- The variance of MR values is higher at low heritability (0.1) but tapers to a smaller band as heritability increases for both models with the exception of Zero-cor model that has 80% of its plants removed when retaining top 5% plants.

- With-cor model shows a gradient improvement in MR compared to Zero-cor model as correlation increase when heritability is between low (0.1) to moderate (0.6). This improvement is no longer present when heritability is higher than 0.6.

- When we cull 80% of plants, the MR value has a distinguishable poor result as compared to other culling percentages independent from heritability in the With-cor model while retaining top 5% plants.

When retaining top 10% of plants:

- The MR curve follows a concave pattern when 80% of the subjects are removed for both models.
- In with-cor models, when correlation between breeding value is high (0.9), the MR remains stable for low to moderate heritability (from 0.1 to 0.6) but then proceeds to follow a downward trend with the exception of removing 80% of plants.
- When 80% of plants were removed we observed the lowest MR when heritability is 0.5 with no correlation between breeding values. The lowest MR did not change much when correlation increases.
- The MR curves follow a downward trend for the other culling percentages.

When retaining top 20% of plants:

- The MR curve follows a concave pattern when 60% and 80% of the subjects are removed for both models with the exception of With-cor model having high correlation (0.9).
- In with-cor models, when correlation is high (0.9), the MR remains stable as heritability increases for all culling percentages.
- In Zero-cor model, When 60% and 80% of subjects were removed, the lowest MR we observed did not change much when correlation varies.
- The MR curves follow a downward trend for the other culling percentages.

When retaining top 30% of plants:

- The MR curve follows a concave pattern when 40%, 60% and 80% of the subjects are removed for both models with the exception of With-cor model having high correlation (0.9).
- In with-cor models, when correlation is high (0.9), the MR remains stable as heritability increases for all culling percentages.

- In Zero-cor model, When 40%, 60% and 80% of subjects were removed the lowest MR we observed did not change much as correlation increases.

- The MR curves follow a downward trend for the other culling percentages.

When retaining top 40% of plants:

- The MR curve follows a shallow concave pattern of all the other culling percentage but 0% for both model.

4.3 Trial Management

In this section, we summarise the results for implementing three trial management strategies with the aim to reduce that culling has on estimating BLUPS. Each strategy was tested with both missing data scenario as as described in section 3.2. Results of plants with one phenotypic observation missing is summarised in subsection 4.3.1. Results of plants missing both of its phenotypic observation is summarised in subsection 4.3.2.

Phenotypic data with partial T_1 observation missing

The idea of trial management strategies is that a certain percentages of plants are saved from the culled population by different sampling methods (randomly, sequentially and systematically). The phenotypic observations of those plants are henceforth no longer missing. Breeding values were estimated using multi-trait LMMs. We ran simulations with 10%, 30% and 50% of the observations missing, for each missing percentage for ran the simulation with retaining 10%, 30% and 50% of the to-be-culled plants. Results of trial management strategies were compared with the results from a reference model, in which individuals were culled using the method described in item 1 while no plants were saved.

Remove 10% phenotypic observations from data set

All three strategies do not show too much improvements compared to the reference model between all the saving and retaining percentages with the exception of retaining top 40% plants while 50% of culled plants are saved. This summary result is present in

Figure 4.5. The x-axis indicates the heritability, the figures on the top of the panel show the top-n percentage of retained plants. The y-axis indicates the MR. The numbers on the right hand side of panel are the true correlation between traits on breeding value level and the numbers on the top panels are the percentages of plants saved. Line colours are used to show the different trial and reference strategies.

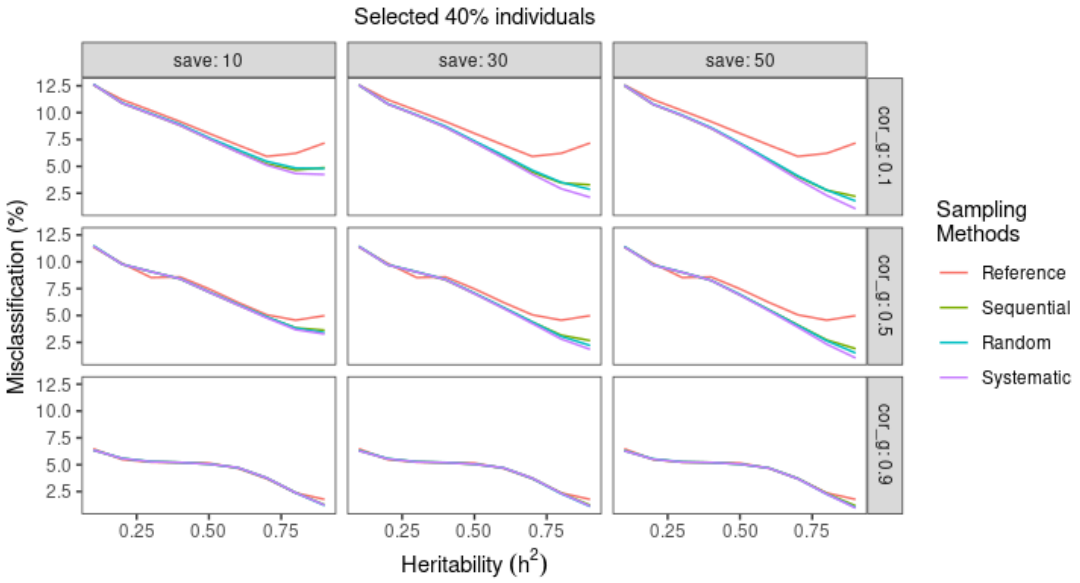


FIGURE 4.5. Misclassification rates of retaining top 40% plants for the three trial management strategies and the reference strategy when 10% of plants are plants.

When retaining top 40% plants:

- All three strategies show an improvement in MR when heritability of trait T_1 is higher than 0.7, 0.75, 0.85, while the correlation between T_1 and T_2 is 0.1, 0.5 and 0.9, respectively, as 10% of the individuals are saved.

- All three strategies show an improvement in MR when heritability of T_1 is higher than 0.3, 0.4, 0.85, while the correlation between T_1 and T_2 is 0.1, 0.5, and 0.9, respectively, as 30% of the individuals are saved.

- All three strategies show an improvement in MR when heritability of T_1 is higher than 0.2, 0.3, 0.85 while correlation between T_1 and T_2 is 0.1, 0.5, and 0.9, respectively, as 50% of the individuals are saved.

- MR has the highest improvement when plants were saved systematically. Saving plants randomly and sequentially also show some improvements. Difference between improvement of three strategies become smaller when correlation between T_1 and T_2 increase.

Remove 30% of phenotypic observations from the data set

All three strategies do not show too much improvement compared to the reference model when retaining top 1%, 5%, 10% and 20% plants while 30% plants are culled. Saving 10% of the plants only shows mild improvement in MR. Then, in Figure 4.6 the result of comparing between the three strategies and the reference model when 30% plants are saved while retaining top 30% plants is presented. The result of the comparison between the three strategies and the reference model when 30% plants are saved while retaining top 40% plants is shown in Figure 4.7.

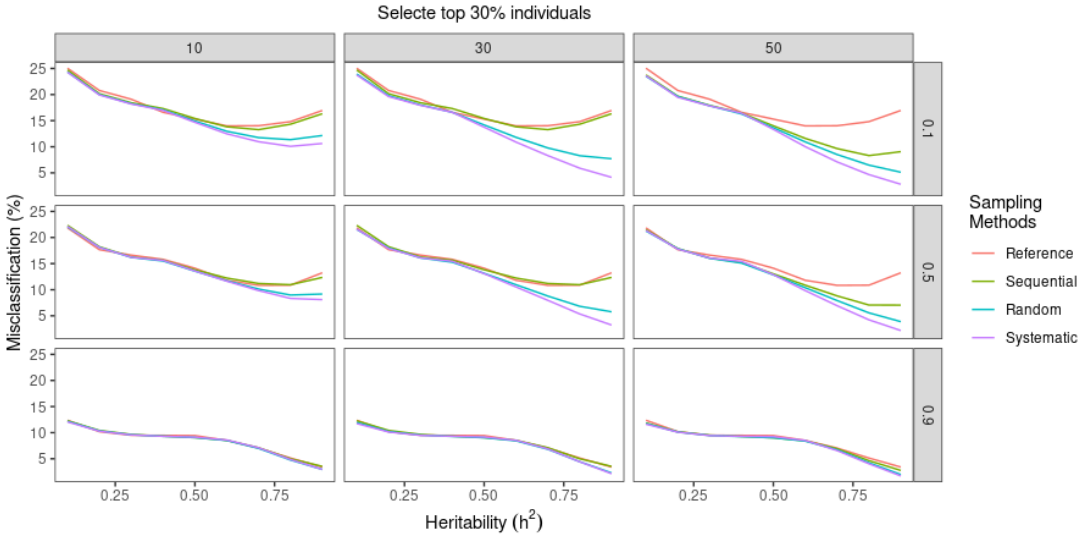


FIGURE 4.6. Misclassification rate of retaining top 30% plants between three trial management strategies and reference model when 30% of plants are culled.

When retaining top 30% plants:

- Saving plants randomly and systematically shows an improvement in MR when the heritability of T_1 is higher than 0.6, 0.7 while correlation between T_1 and T_2 is 0.1, 0.5 as 10% plants are saved. None of the strategies shows an improvement when the correlation is 0.9.

- Saving plants randomly and systematically shows an improvement in MR when the heritability of T_1 is higher than 0.4, 0.5 and 0.8 while the correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 as 30% plants saved.

- All strategies show an improvement in MR when heritability of T_1 is higher than 0.3, 0.4 and 0.7 while correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 as 50% of the plants are saved.

- Saving plants randomly and systematically improves the MR for all the saving percentages. MR has the highest improvement when plants are saved systematically. Saving plants randomly also shows some improvement. The difference between improvement for the two strategies becomes smaller when the correlation between T_1 and T_2 increases. Saving plants sequentially appears to improve only the MR when 50% of the plants are saved, and it shows the least improvement among all three strategies.

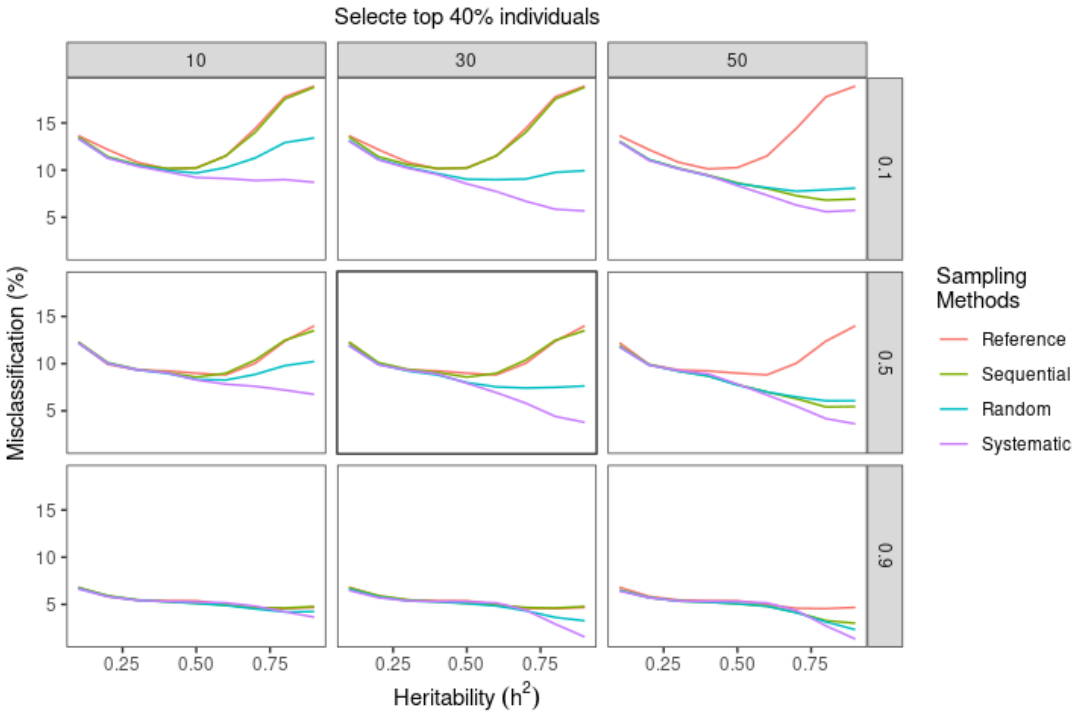


FIGURE 4.7. Misclassification rate of retaining top 30% plants between three trial management strategies and reference model when 40% culled population were saved.

When retaining the top 40% of the plants:

- Saving plants randomly and systematically shows an improvement in MR when the heritability of T_1 is higher than 0.4, 0.5 and 0.75 while the correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 as 10% of the plants are saved.

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher than 0.35, 0.4 and 0.7 while the correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 as 30% of the plants are saved.

- All strategies show an improvement in MR when the heritability of T_1 is higher than 0.1, 0.35 and 0.7 while the correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 as 50% of the plants are saved.

- Saving plants randomly and systematically improves the MR for all the saving percentages. MR has the highest improvement when plants are saved systematically. Saving plants randomly also shows some improvements. The difference between improvement for the two strategies becomes smaller when the correlation between T_1 and T_2 increases. Saving plants sequentially appears to improve only the MR when we save 50% of the plants, and it shows a slightly better improvement than for the random strategy.

Remove 50% phenotypic observations from the data set

All three strategies do not show too much improvement compares to the reference model when retaining the top 1%, 5% and 10% of plants while 50% plants are saved. Then, Figure 4.8 presents the result of the comparison between the three strategies and the reference model when retaining the top 20% of plants. Figure 4.9 shows the result of the comparison between the three strategies and the reference model when retaining the top 30% of plants. The result for the comparison between the three strategies and the reference model when retaining the top 40% of plants is presented in Figure 4.10.

When retaining top 20% of plants:

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher than 0.5, 0.6 while correlation between T_1 and T_2 is

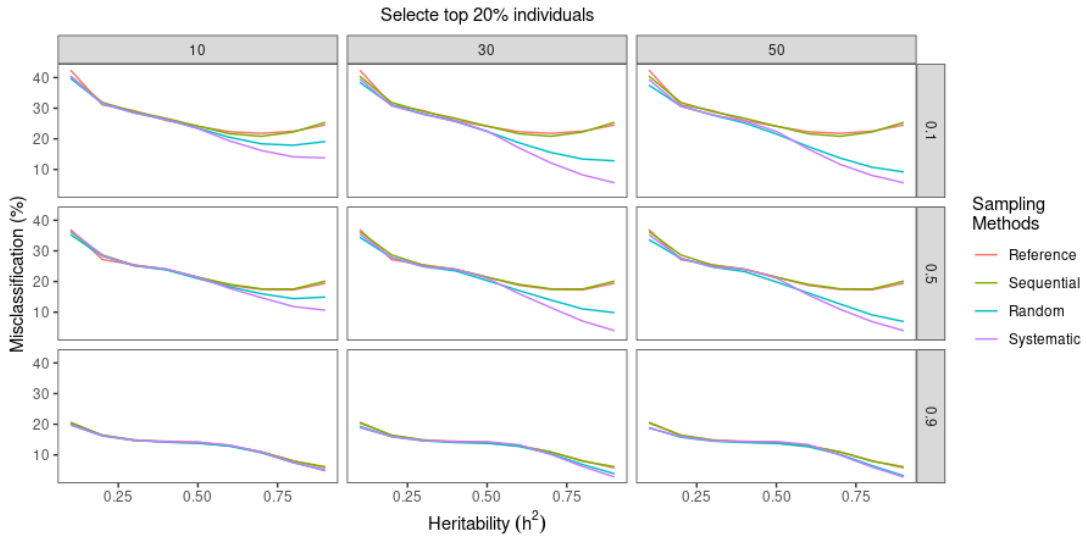


FIGURE 4.8. Misclassification rate of retaining top 20% plants between three trial management strategies and reference model when 50% plants culled.

0.1, 0.5 as 10% plants are saved. None of the strategies shows an improvement when the correlation is 0.9.

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher than 0.4, 0.5 and 0.7 while correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 as 30% of plants are saved.

- Save plants randomly and systematically showed an improvement in MR when heritability of T_1 is higher than 0.4, 0.55 and 0.65 while correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 as 50% of the plants are saved.

- Saving plants randomly and systematically improves the MR in all the saving percentages. MR has the highest improvement when plants are saved systematically. Saving plants randomly also shows some improvement. The difference in improvement for the two strategies become smaller when the correlation between T_1 and T_2 increases. Saving plants sequentially does not appear to be helpful in this combination of culling and retaining percentage.

When retaining top 30% of plants:

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher than 0.5, 0.6 while correlation between T_1 and T_2 is 0.1,

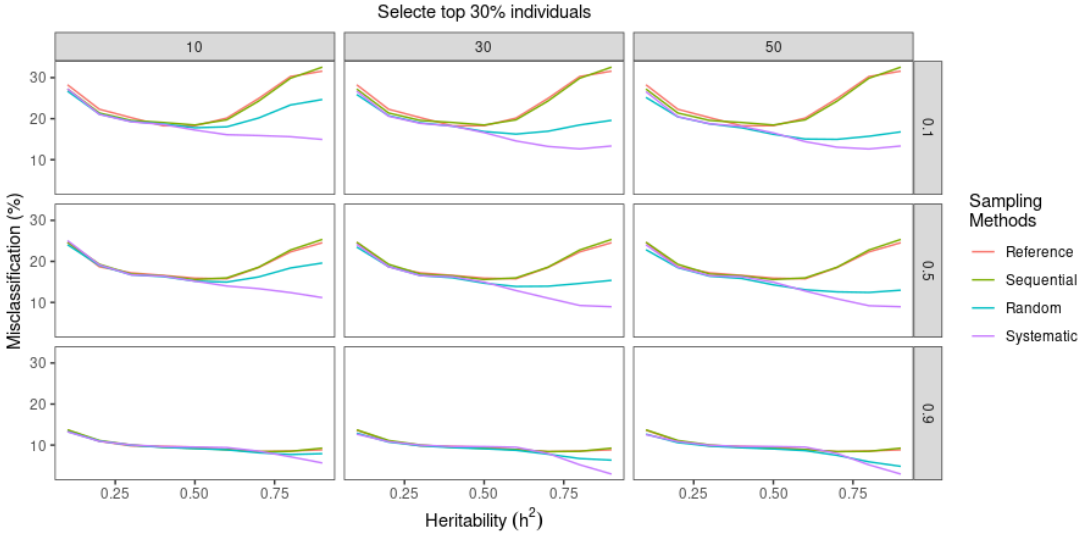


FIGURE 4.9. Misclassification rate of retaining top 30% plants between three trial management strategies and reference model when 50% plants culled.

0.5 as 10% of the plants are saved. None of the strategies shows an improvement when the correlation is 0.9.

- Saveing plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher than 0.4, 0.5 and 0.7 while correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 as 30% of plants are saved.

- Saveing plants randomly and systematically showed an improvement in MR when heritability of T_1 is higher than 0.4, 0.55 and 0.65 while correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 as 50% of plants are saved.

- Saving plants randomly and systematically improves MR for all the saving percentages. MR has the highest improvement when plants are saved systematically. Saving plants randomly also shows some improvements. The difference between improvement for the two strategies become smaller when correlation between T_1 and T_2 increases. Saving plants sequentially does not appear to be helpful in this combination of culling and retaining percentage.

When retaining top 40% of plants:

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher than 0.4, 0.5 and 0.7 while correlation between T_1 and

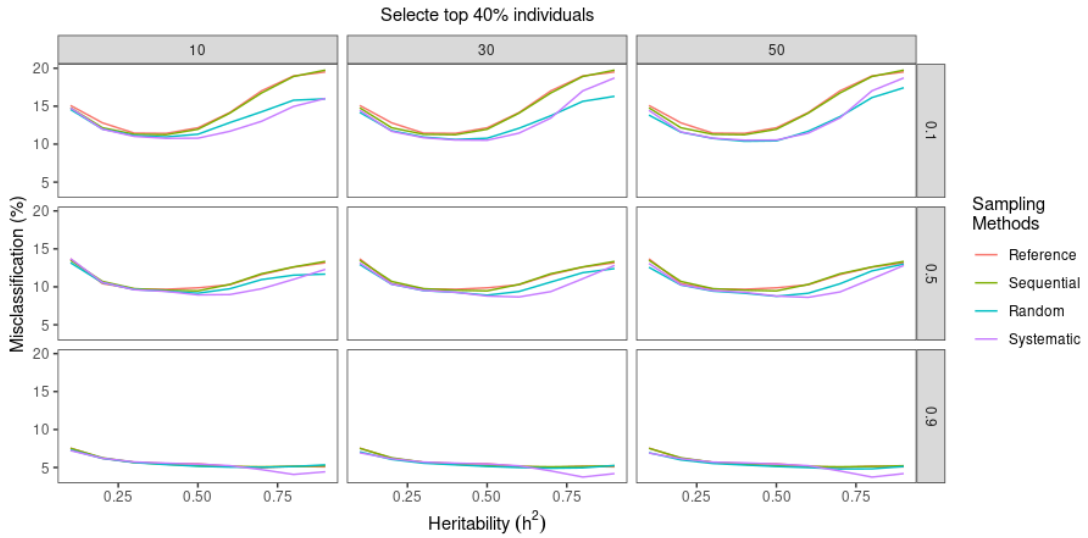


FIGURE 4.10. Misclassification rate of retaining top 40% plants between three trial management strategies and reference model when 50% plants culled.

T_2 is 0.1, 0.5 and 0.9 as 10% of plants are saved.

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher than 0.3, 0.4 and 0.65 while correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 as 30% of plants are saved.

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher than 0.1, 0.3 and 0.65 while correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 as 50% of plants are saved.

- Saving plants randomly and systematically improves the MR in all the saving percentages. MR has the highest improvement when plants were saved systematically. Saving plants randomly also show some improvements. The difference between improvement for the two strategies become smaller when correlation between T_1 and T_2 increase. Saving plants sequential does not appears to be helpful in this combination of culling and retaining percentage.

Plants with both phenotypic data missing

In this section, simulations carried out to test the trial management strategies when culling remove a plant completely (both phenotypic observations are missing). Result from data which missing 10% plants is summarised in section 4.3.2 plants, result of

missing 30% plants is summarised in **??** and result of missing 50% plants is summarised in **??**. For each missing percentage, three sampling strategies were simulated to save 10%, 30% and 50% plants from the culled population.

Remove 10% **of plant from the data**

All three strategies do not show too much improvement compares to the reference model when retaining top 1%, 5%, 10%, 20% and 30% of plants while 10% of plants are culled. Figure 4.11 presents result of comparison between the three strategies and reference model when retaining top 40% plants.

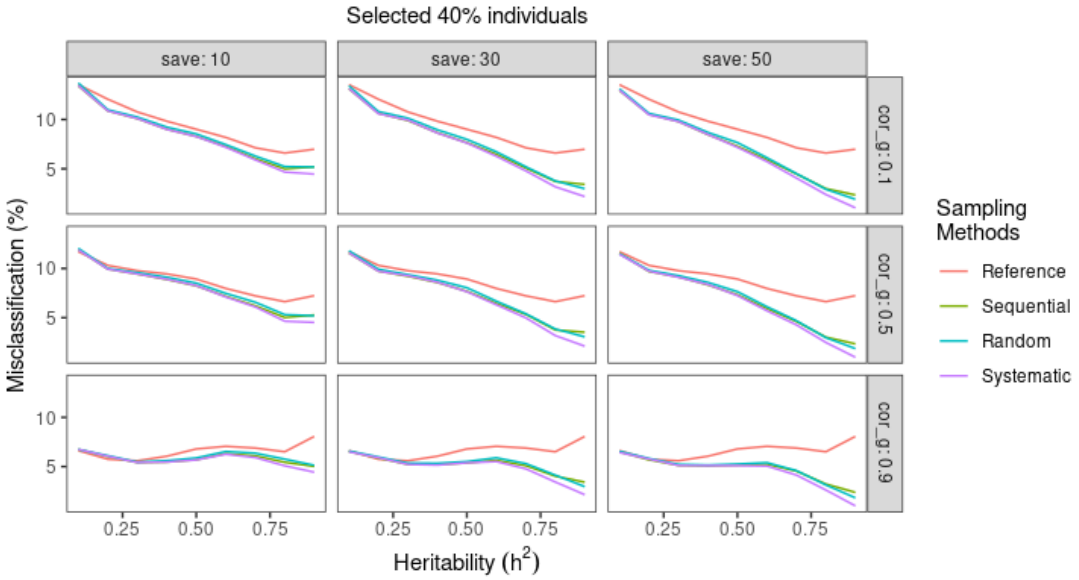


FIGURE 4.11. Misclassification rate between three trial management strategies when retaining top 40% plants while 10% plants culled.

When retaining top 40% of plants:

- All three strategies show an improvement in MR when heritability of T_1 is higher than 0.1, 0.2, 0.3 while correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 respectively as 10% of plans are saved.

- All three strategies show an improvement in MR when heritability of T_1 is higher than 0.1, 0.15, 0.3 while correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 respectively as 30% of plants are saved.

- All three strategies show an improvement in MR when heritability of T_1 is higher than 0.1, 0.1, 0.25 while correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 respectively as 50% of plants are saved.

- MR has the highest improvement when plants are saved systematically. Saving plants randomly and sequentially also shows some improvements. There is an increasing improvement in MR while more plants were saved.

Remove 30% of plant from the data

All three strategies do not show too much improvement compares to the reference model when retaining top 1%, 5%, 10% and 20% plants while 30% plants are culled. So Figure 4.12 presents result of comparison between the three strategies and reference model when retaining top 30% plants. Figure 4.13 presents result of comparison between the three strategies and reference model when retaining top 40% plants.

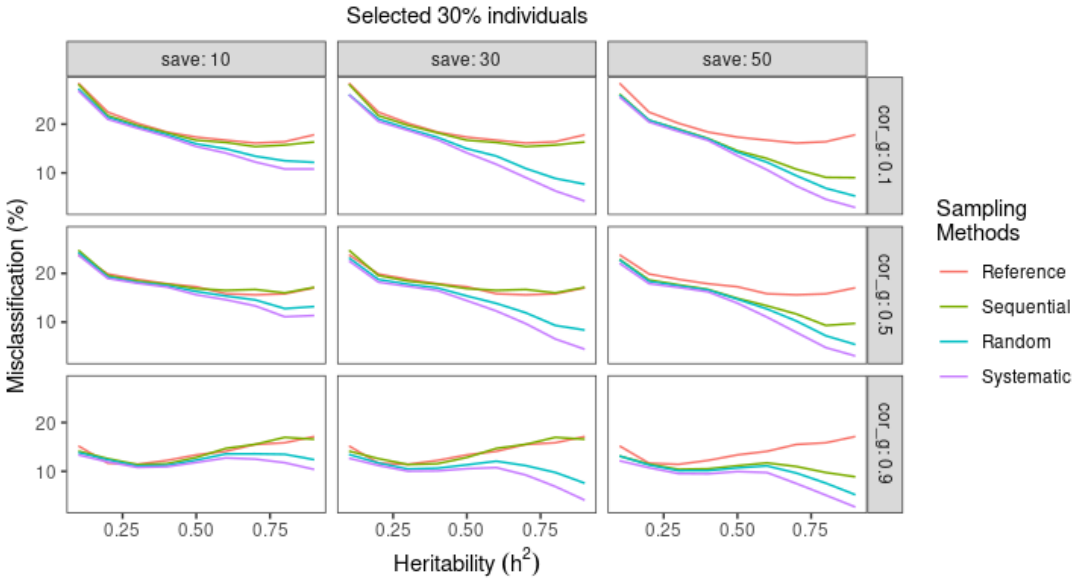


FIGURE 4.12. Misclassification rate between three trial management strategies when retaining top 30% plants while 30% plants culled.

When retain top 30% of plants:

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher than 0.4, 0.5 and 0.55 while correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 as 10% of plants are saved.

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher 0.1 for all correlation (0.1, 0.5 and 0.9) T_1 and T_2 as 30% of plants are saved.

- All strategies shows an improvement in MR when heritability of T_1 is higher than 0.1 for all correlation (0.1, 0.5 and 0.9) between T_1 and T_2 as 50% of plants are saved.

- Saving plants randomly and systematically improve MR in all the saving percentages. MR has the highest improvement when plants were saved systematically. Saving plants randomly also shows some improvement. The difference in improvement for the two strategies becomes smaller when correlation between T_1 and T_2 increases. Saving plants sequential appears to improve MR only when 50% of plants are saved, and it shows the least improvement among all three strategies. There is positive relationship between improvement in MR and amount of plants saved.

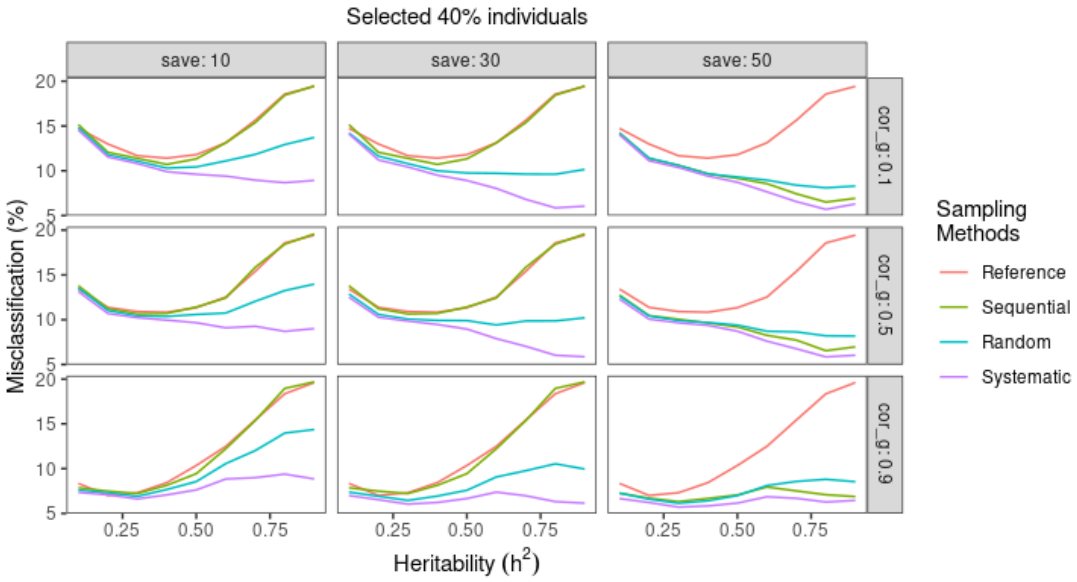


FIGURE 4.13. Misclassification rate between three trial management strategies when retaining top 40% of plants while 30% plants culled.

when retain top 40% plants:

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher 0.1 for all correlation (0.1, 0.5 and 0.9) between T_1 and T_2 when 10% of plants are saved.

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher 0.1 for all correlation (0.1, 0.5 and 0.9) between T_1 and T_2 when 30% of plants are saved.

- All strategies show an improvement in MR when heritability of T_1 is higher than 0.1 for all correlation (0.1, 0.5 and 0.9) between T_1 and T_2 when 50% of plants are saved.

- Saving plants randomly and systematically improve MR for all the saving percentages. MR has the highest improvement when plants are saved systematically. Saving plants randomly also shows some improvements. The difference in improvement for the two strategies become smaller when correlation between T_1 and T_2 increases. Saving plants sequentially appears to improve MR only when 50% of plants are saved, and it shows a slightly better improvement compares to random strategy. There is positive relationship between improvement in MR and amount of plants are saved.

Remove 50% of plant from the data

All three strategies do not show too much improvement compares to the reference model when retaining top 1%, 5% and 10% plants when 50% plants are saved. Figure 4.14 presents result of comparison between the three strategies and reference model when retaining top 20% plants. Figure 4.15 presents result of comparison between the three strategies and reference model when retaining top 30% plants. Result of comparison between the three strategies and reference model top 40% plants retain is present in Figure 4.16

When retaining top 20% of plants:

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher than 0.4, 0.5 and 0.6 while correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 respectively as 10% of plants are saved.

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher than 0.3, 0.2 and 0.1 while correlation between T_1 and

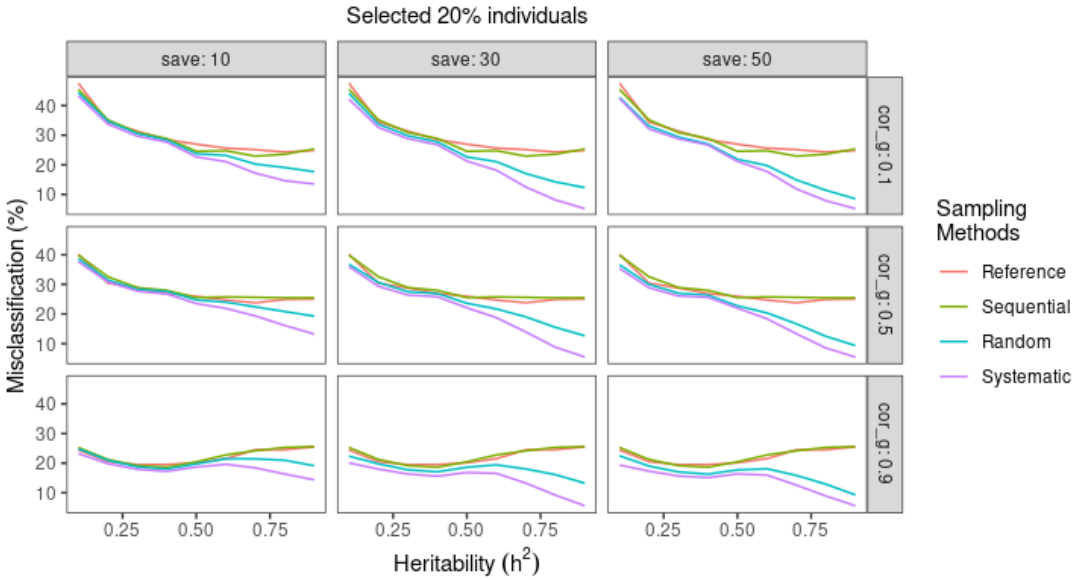


FIGURE 4.14. Misclassification rate of retaining top 20% plants between three trial management strategies and reference model when 50% plants culled.

T_2 is 0.1, 0.5 and 0.9 as 30% of plants are saved.

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher than 0,1 for all correlation (0.1, 0.5 and 0.9) between T_1 and T_2 as 50% of plants are saved.

- Saving plants randomly and systematically improves MR in all the saving percentages. MR has the highest improvement when plants are saved systematically. Saving plants randomly also shows some improvements. The difference between improvement of the two strategies become smaller when correlation between T_1 and T_2 increases. Saving plants sequentially does not appear to be helpful in this combination of culling and retaining percentage.

When retaining top 30% of plants:

- Saving plants randomly and systematically showes an improvement in MR when heritability of T_1 is higher for all correlation (0.1, 0.5, 0.9) between T_1 and T_2 when 10% of plants are saved.

- Saving plants randomly and systematically showes an improvement in MR when heritability of T_1 is higher for all correlation (0.1, 0.5, 0.9) between T_1 and T_2 when 30% of plants are saved.

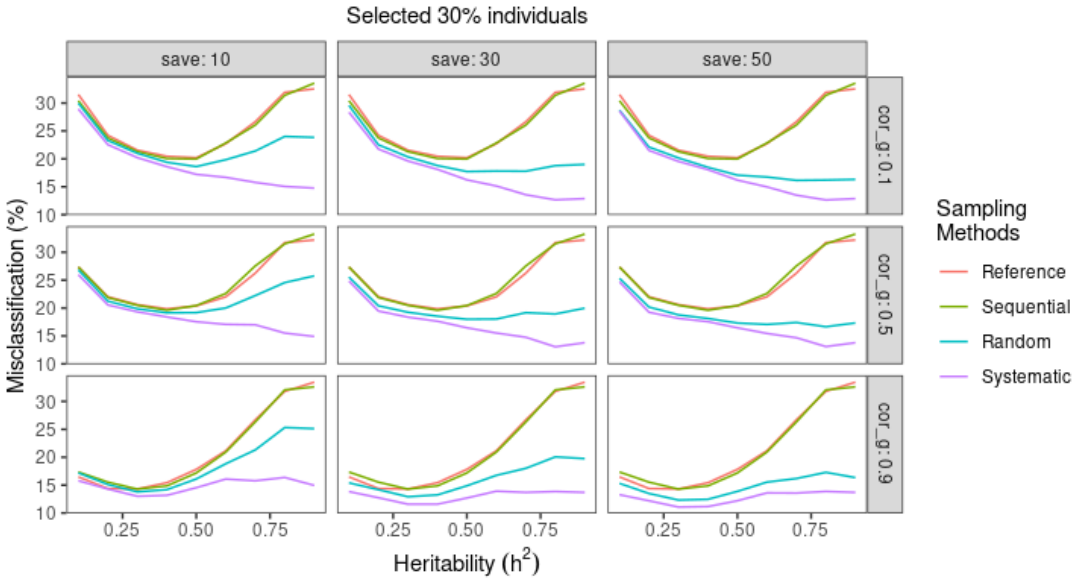


FIGURE 4.15. Misclassification rate of retaining top 30% plants between three trial management strategies and reference model when 50% plants culled.

- Saving plants randomly and systematically showes an improvement in MR when heritability of T_1 is higher for all correlation (0.1, 0.5, 0.9) between T_1 and T_2 when 50% of plants are saved.

- Saving plants randomly and systematically improves MR in all the saving percentages. when saving 10% of plants, MR has the highest improvement when plants are saved systematically, saving plants randomly also shows some improvements. When save 30% and 50% of plants, save plants randomly appear to improve MR more when heritability is high. Saving plants sequentially does not appear helpful in this combination of culling and retaining percentage.

When retaining top 40% of plants:

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher than 0.4, 0.5 and 0.7 while correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 as 10% of plants are saved.

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher than 0.3, 0.4 and 0.65 while correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 as 30% of plants are saved.

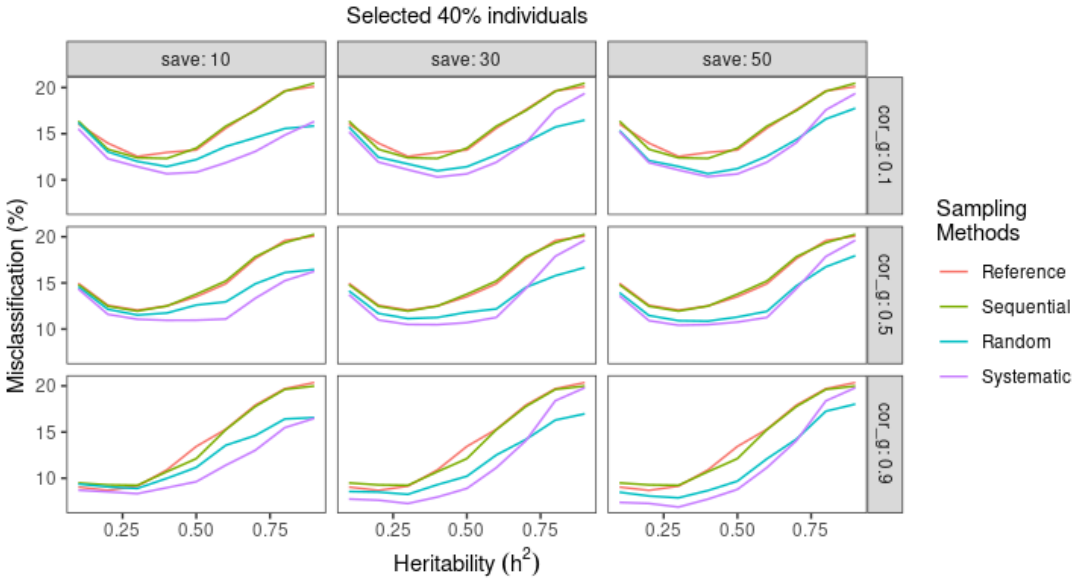


FIGURE 4.16. Misclassification rate of retaining top 40% plants between three trial management strategies and reference model when 50% plants culled.

- Saving plants randomly and systematically shows an improvement in MR when heritability of T_1 is higher than 0.1, 0.3 and 0.65 while correlation between T_1 and T_2 is 0.1, 0.5 and 0.9 as 50% of plants saved.

- Saving plants randomly and systematically improves MR in all the saving percentages. MR has the highest improvement when plants were saved systematically. Saving plants randomly also shows some improvements. The difference in improvement for the two strategies becomes smaller when correlation between T_1 and T_2 increases. Saving plants sequentially does not appear to be helpful in this combination of culling and retaining percentage.

Chapter 5

Discussion & Conclusion

5.1 Effect of culling on breeding values estimation

The effect of culling on breeding values is quantified by using simulated data in terms of the consistency of genotype rankings. Combinations of various culling and selection percentage are applied on the data to mimic the real process.

As expected, culling affects the estimation of additive genetic effects and consequently the ranking of genotypes. When retaining a small number of plants, MR and heritability has a negative linear relationship and it is independent from the culling percentage. The difference of MR between culling percentage is discernible when heritability is high and becomes marginal as heritability decreases. When heritability is low, the variation of a trait contributes mostly by non-genetic factors, in other words, the environmental effect. When a trait is influenced largely by its environment, there is higher proportion of noise in the data, which makes it harder filter out the effect of each genotype. Therefore, even when all plants are retained, MR is still high due to the unstable phenotypic performance. As the heritability increases, variation of the trait become influenced by its genetic effect. This makes it easier for LMMs to separate additive genetic effects from environmental effects. Because of this the MR is low when high percentages of plants were culled.

Unexpectedly, in some situation the lowest MR no longer appears for the highest heritability as the relationship between MR and heritability becomes quadratic. Some MR curves displayed a concave pattern as heritability increased. We believe that the

combination of heritability and the number of plants retained are the reason why some of these curves have such shapes.

When a trait has low heritability, variation between genotypes is high due to a high proportion of environmental effects, which causes breeding values to have a wide distribution. This larger uncertainty about predicting breeding values increases the chance of phenotypic data distributions for different genotypes to overlap. As heritability increases, variation within genotype become smaller owing to a decrease of environmental effect. This smaller variation decreases the spread of each genotype distribution, which results in a stronger separation between the phenotypic data distributions of different genotype. According to this theory, when culling occurs sequentially, the removal of all observations from the same genotype is more likely for a trait with high heritability compared to a trait with low heritability.

For example, Figure 5.1 shows the distribution plots of two different genotypes (Genotype1 and Genotype2) given a certain heritability for the trait. All subjects from the same genotype can be sampled from the genotypes distribution. The heritabilities are 0.1 in Figure 5.1(A), 0.5 in Figure 5.1(B) and 0.9 in Figure 5.1(C). If a trait has low heritability, the distribution of breeding values for that trait will be more widely spread and the distribution curves for different genotypes can have a larger overlap compared to those with a trait that has high heritability. This is due to the high ratio of noise in the data for less heritable traits. Individuals sampled from distributions in Figure 5.1(A) are more scattered. As the heritability increases, the distribution curves are less stretched and chances of an overlap of distributions reduces. The overlap between the distributions in Figure 5.1(B) is less, compared with Figure 5.1(A). In Figure 5.1(C) the heritability is 0.9 and there is barely any overlapping between genotypes.

When we mimic the culling process, subjects are removed according to their phenotypic observations, in our case, in ascending order. Since the phenotypic data is more scattered when the heritability is low, it is unlikely to have both subjects from the same genotype removed. Breeding values can be estimated using the observation of the remaining plants. This is why the MR initially decreases when the heritability is low. If plants are culled according to traits with high heritability, phenotypic observation

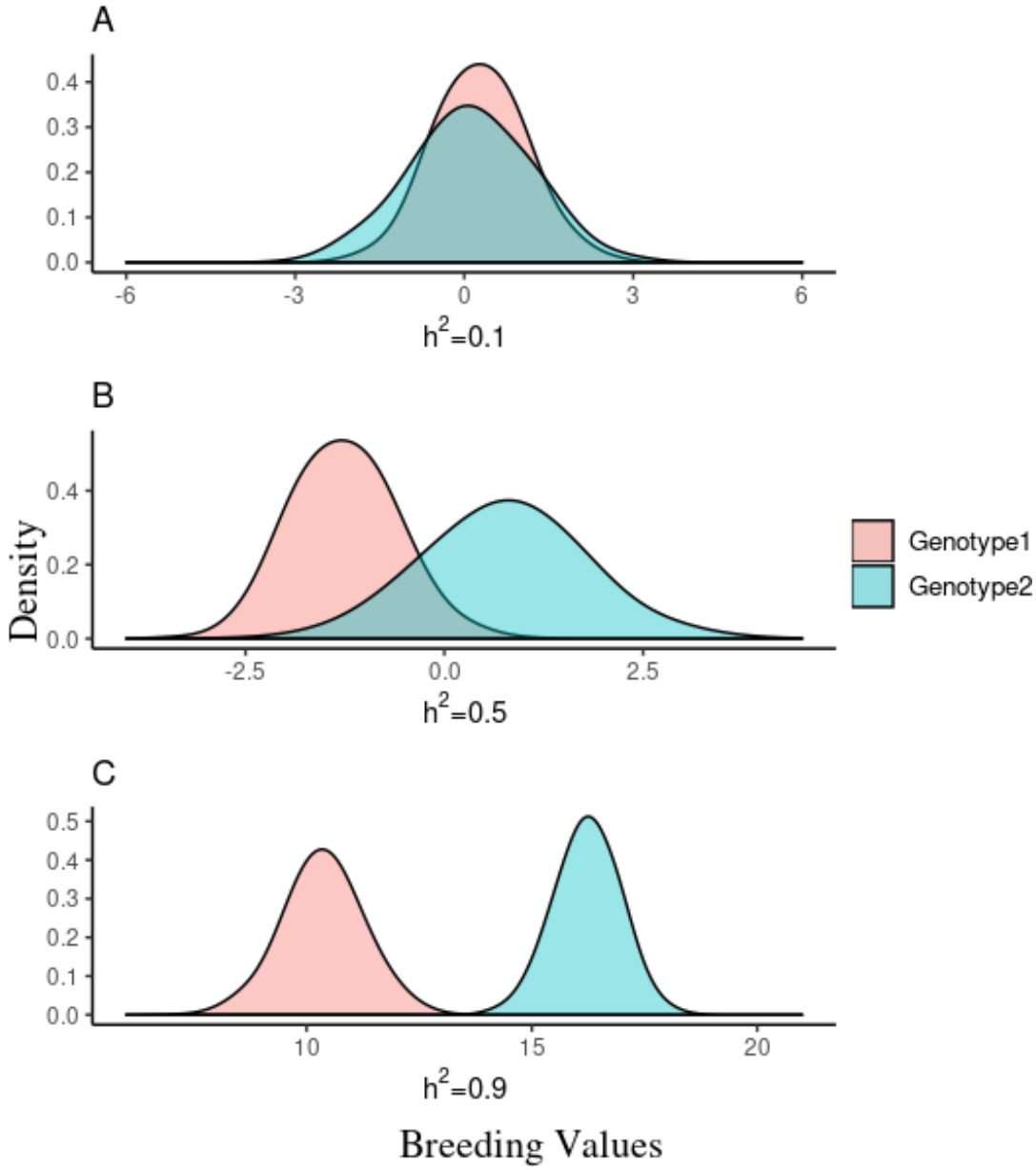


FIGURE 5.1. Distributions of breeding value with different heritability

from the same genotype are highly clustered. It is more likely that all plants from the same genotype are being removed. If a genotype has no phenotypic information left due to the culling process, our best estimates for that genotype is the population average, given the pedigree structure and correlation with other traits.

The number of plants retained also plays a role. When retaining a small number of plants, MRs are less likely to be influenced by culling. Retaining plants that have less heritable traits increases the chance of having at least one plant left in the data due to the scatter property of the distribution. If retaining plants according to traits with high

heritability, then it is likely that the plants retained are both from the same genotype. It is the one that has higher mean phenotype. However, when retaining a large number of plants according to trait with high heritability, it is likely that we select a genotype with no phenotypic data available. The population average is a less accurate prediction of breeding values compares to those that were predicted from the same genotype. The combination of high heritability and high number of retained plants is the reasons why MR increased again after a downward trend.

The number of genotypes that remain in the data after culling further helps us elaborate on the theory and can strengthen our interpretation. Figure 5.2 shows the average number of genotypes remaining in the data after culling for different values of heritability. For example, if all plants of the same genotype are culled, then that genotype is no longer present in the data. If all plants but one are culled the genotype is still present in the data. The x-axis shows the heritability of a trait and the y-axis is the average number of genotypes remaining in a data after culling. Line colours are used to show the percentage of plants removed. When culling is based on a highly heritable trait, the phenotypic observation are less scattered hence, the number of genotypes left in the data is less compared to with low heritability. Regardless of the percentage of subjects removed, the number of genotypes in the data shows a downward trend as the heritability increases. For example, the purple line suggests that when removing 80% of subjects from the data, number of genotypes left in the data is about 440 when heritability is 0.25. There were only 310 genotypes left when heritability is 0.75 given the same number of subjects removed.

Overall, our simulation study shows that in breeding trials when each genotype has more than one plant, it is unlikely to miss out on top-ranking cultivars when the number of selections is small while culling occurred sequentially on traits with high heritability. When retaining more plants, there is an increasing risk of mistakenly selecting a low-ranking cultivar at high culling percentage. The relation between heritability and MR is no longer linear as the lowest MR shifted to smaller heritability value while culling percentage and selection percentages increased.

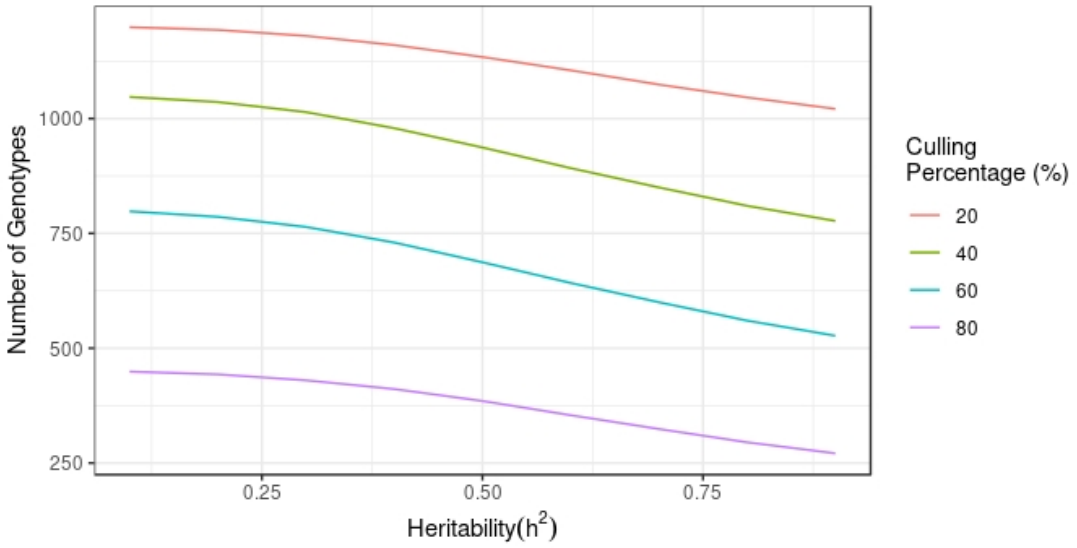


FIGURE 5.2. Average number of genotypes left after selection

5.2 Multi-trait model approach of BLUP

Since culling introduces a bias into breeding values estimation, and it is inevitable in breeding trials, we evaluated if a multi-trait LMM has the potential to improve breeding value estimation accuracy. The idea is that by including phenotypic observation of other correlated traits and (co)variance between traits, breeding value estimation would be less biased. From the results of our simulation study, multi-trait LMM generally lead to much higher accuracy of breeding values prediction when heritability is low ($h^2 \leq 0.1$) than a single-trait LMM. This was expected because the improvement of breeding value estimation accuracy is more profound for a trait with low heritability than a higher one (Volpato et al., 2019). Jia and Jannink (2012) compared the prediction accuracy for a low (0.1), medium (0.5) and high heritability trait (0.8). The study showed that in a multi-trait model, prediction accuracy increased for low heritability trait while no improvement was observed for medium and high heritability trait.

Another factor influencing breeding value estimation accuracy is the correlation between genetic effects and correlation between environmental effects. In our simulation, the correlation between non-genetic effect is a fixed value (0.9), so we are only going to discuss the correlation between genetic effect. Three levels of correlation (low, moderate and high) between genetic effect were simulated. The result shows that when the correlation between genetic effect was small, the difference of breeding value estimation

accuracy between single-trait model and multi-trait model is small and independent from culling percentage. The difference between the two models increased as correlation increases. When the correlation between traits is high, the accuracy of prediction drastically improves. This is because when the correlation between genetic effect is small, the extra information gained from a correlated trait is small and therefore, the differences in breeding value estimation accuracy increases when correlation increases (Koerhuis and van der Werf, 1994).

An unexpected result from the study was when culling caused both phenotypic observations missing, we expect to see no improvement in MR as correlation increased because no extra information was gained from the correlated trait. However, from the results, we can still see some improvement in MR using a multi-trait LMM. When a plant has no phenotypic information available, the estimated breeding value is equal to the population mean, given the pedigree structure and correlation with other traits. When traits have high correlation between breeding values, the true breeding values are concentrated but are scattered when they have low correlation. When given the population mean as breeding value estimation, the distance between the true breeding value and estimated breeding value is smaller when correlation is high than when correlation is low. This is why even when a trait has no phenotypic data available, a highly correlated trait can be used as reference to estimate its breeding value.

Overall, MR calculated from results of With-cor model are lower than results of Zero-cor model given the same culling percentage and heritability, and as correlation increase the differences became more significant. As more subjects retained, differences between MR calculated from the two models are gradually decrease. In practise, when traits has missing phenotypic information due to cull, and breeders wants to select genotype using the incomplete data set, correlated traits can be used in a multi-trait model to compensate the missing information. This method is even helpful when plants has no phenotypic observation available, the correlation between remaining data is still helpful in breeding value estimation.

5.3 Trial management

Three sampling strategies of saving a number of plants from the culled population have been evaluated using a simulation study. The aim was to evaluate the effect of each sampling strategy on the the accuracy of breeding values estimation. The three sampling method are: sampling randomly, sequentially and systematically. The effect of each strategy depends on the heritability of trait of interest, percentages of plants saved, selected, the correlation between traits of interest and other recorded traits. We will discuss the effect of applying trial management for each of these variables.

Systematically saving plants proved to be the most efficient method overall. Saving plants randomly also smoothed out the concave MR curves previously described, but not as much as saving plants systematically. Saving plants sequentially has the poorest performance out of the three strategies on average. The accuracy of breeding value estimation does not improve at all, in some cases MR are even higher compared with not saving any no plants at all. hence from here on we will only further discuss the random and systematic sampling strategies.

As mentioned in section 5.1, the MRs have a negative linear relationship with the heritability when a small amount of plants were selected. The MR decrease when the heritability of a target trait increases. Given an increased number of selected plants, the relationship is no longer linear. Some trial management strategies are efficient, but only in terms of smoothing the concave shapes, not decrease the slop of MR curve. Based on our simulation, trial management strategies do not improve the MR when (1) heritability of the trait of interest is lower than 0.5.(2) amount of selected plants are small (less than 10%). Relationship between MR and heritability remains linear when the heritability is low and the selection amount is small, hence naturally none of the strategies improves the accuracy of breeding values prediction compares to the reference model.

Effect of strategies vary by the amount of plants saved, none of the strategy improved breeding value accuracy by saving 10% plants, save 30% plants shows continuous improvement as more plants selected, however, this improvement is weakened when 50%

plants were saved. Last, strategies also performed differently according to the correlation between the trait of interest and other traits used in the model, improvement brought by different strategies diverged more from each other when the correlation is small, improvement difference between random and systematic strategy become smaller when correlation increased. But this divergent between strategies only appeared when a plant is missing partial observation.

Saving plants sequentially does not improve the MR when selecting smaller amount of plants is because it was only calculated for selected plants. These plants are those carrying genotype which ranked at the top, saving plants from the lower tail of the phenotype distribution is unlikely going to provide useful information to improve MR. In Figure 5.1(C), when a trait has high heritability, then phenotype distribution of each genotype were separate without any overlapping. Plants were culled according to its phenotype in ascending order. The separation between distributions make the selection and culling focused on different distribution. Selected plants are mostly from Genotype2 and culled plants were from Genotype1, so all strategies we are proposing can help save missing information from Genotype1, random and systematic strategies saved more information from different genotypes than sequential method. The same reason applies to moderate heritability, Figure 5.1(B) has some overlap between distributions, some missing information can be offset by saved plants. When heritability is low, majority areas under distribution curves are overlapped as Figure 5.1(A) showed. It is less likely to have both plants removed from the same genotypes, so when the trait of interest has low heritability, it does not benefit from saving strategies we proposed.

For similar reason as why selecting a small number of top-ranking plants does not benefit from the strategies is because MR curve had a linear downward trend when select small percentage of plants, strategies of saving plants helped to smooth the concave MR curve, select less than 10% plants does not yet caused the curve to concave, when there is no concave shape in the curve, MR does not benefit from the method we proposed.

Data points of phenotypic observations are much more scattered given a weak correlation between traits of interest and other traits, phenotype from two plant which belong

to the same genotype has a higher chance to be far away from each other. Compares to traits with strong correlation, save plants randomly from a much concentrated scatter plot is more likely to save information from a genotype which has no observation left due to the culling process. This improvement of random strategy only works when a partial observation is missing from a plant, when both observations were gone, then the correlation is no longer a useful factor as at least one observation is needed in order to benefit from the correlation.

Overall, saving plants from culled population can help improving the accuracy of estimate breeding value. Efficiency of this strategies depends on how plants were saved and how much plants going to be select at the end of the trial. Saving plants systematically is the most efficient at all situation, randomly save plants also show improvement, sampling plants from the sequentially can be helpful but only when large number of plant were selected while culled plant are below certain threshold. When culling result in missing information only on the trait of interest, correlation between traits increase the efficiency of some sampling strategies, it seems to work the best when correlation between traits are high. The benefit of correlation does not presence in the scenario when a plant has all of its phenotype missing.

5.4 Conclusion

This thesis investigated and summarised the effect of culling and verified some potential solutions of reducing such bias in both data analysis and data collection. Our study showed that culling introduces bias into the breeding value prediction. This bias affects traits more with low heritability. We also discovered that in clone breeding trials, when retaining a high amount of plants, a high culling percentage increases the chance of missing the genotypes, which have high true breeding values. Applying a multi-trait mixed model with a covariance structure for the genetic effects, helps to reduce the bias in breeding value estimation, which is introduced by culling. Improvements made by such a model works in both situations where culling causes missing information on the genotype of interest or when a genotype is missing all of its phenotypes. Apart from changing the way of analysing data with missing information, saving plants in the field could also reduce the bias introduced by culling.

The culling bias investigated in this study is focus in clone trials, where all the genotypes are represented by more than one plant. We also tried to extend this study to the first stage trial where each genotype is only represented by a single plant. Unfortunately, we found that ASReml is having difficulties in terms of model convergence. We also simplified the data generation process by ignoring some factors in practise, such as, correlation between environmental effect, interaction between genetic effect and environmental effect and spatial effect between plants in different locations, etc.. The multi-trait model used in this study contained only two traits, in reality, more phenotypes from different traits were collect in the field, correlation between more than two traits could make the interpretation of results quite complicated. The trial management strategies we simulated indeed helps to reduce the culling bias, however, there are more sampling methods we could apply, which may help reduce the bias even further (e.g. systematic sampling, breaking the phenotype data into even intervals and sampling from each interval). Also, the reduction of bias by saving plants is not a strategy which can be applied in all situations, for example, when plants were culled due to contagious diseases, saving plants is certainly not be an option.

Bibliography

Appel, L., Strandberg, E., Danell, B., and Lundeheim, N. (1995). Missing data due to culling of pigs before testing and the effects on the genetic evaluation. *Acta Agriculturae Scandinavica A-Animal Sciences*, 45(4):218–227.

Appel, L., Strandberg, E., Danell, B., and Lundeheim, N. (1998). Adjusting for missing data due to culling before testing in genetic evaluations of swine. *Journal of animal science*, 76:1794–802.

Appel, L., Strandberg, E., Danell, B., and Lundeheim, N. (1999). Culling before testing in swine: Identification of culling strategy and estimation of culling precision. *Journal of animal science*, 77:1666–78.

Bauer, A. M., Hoti, F., Reetz, T. C., Schuh, W.-D., Léon, J., and Sillanpää, M. J. (2009). Bayesian prediction of breeding values by accounting for genotype-by-environment interaction in self-pollinating crops. *Genetics Research*, 91(3):193–207.

Ceballos, H., Pérez, J. C., Joaqui Barandica, O., Lenis, J. I., Morante, N., Calle, F., Pino, L., and Hershey, C. H. (2016). Cassava breeding i: The value of breeding value. *Frontiers in Plant Science*, 7:1227.

Corbeil, R. R. and Searle, S. R. (1976). Restricted maximum likelihood (reml) estimation of variance components in the mixed model. *Technometrics*, 18(1):31–38.

Duvick, D. (2005). The contribution of breeding to yield advances in maize (zea mays l.). *Advances in Agronomy*, 86:83—145.

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to quantitative genetics*. Burnt Mill, England : Longman, 4th ed edition.

Faria, C. U. d., Magnabosco, C. d. U., Reyes, A. d. l., Lôbo, R. B., Bezerra, L. A. F., and Sainz, R. D. (2007). Bayesian inference in a quantitative genetic study of growth traits in nelore cattle (bos indicus). *Genetics and Molecular Biology*, 30(3):545–551.

Fisher, R. A. (1919). The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433.

Gianola, D., Fernando, R., and Im, S. (1989). Likelihood inferences in animal breeding under selection: a missing-data theory view point. *Genetics Selection Evolution*, 21:399–414.

Gilmour, A., R.B.J., G., Cullis, B., and Thompson, R. (2009). Asreml user guide release 3.0. *VSN International Ltd, Hemel Hemptead, HP1 1ES, UK*.

Gurka, M. (2006). Selecting the best linear mixed model under reml. *The American Statistician*, 60(1):19–26.

Hallauer, A. R. (2011). Evolution of plant breeding. *Crop Breeding and Applied Biotechnology*, 11:197–206.

Harris, B. and Johnson, D. (2010). Genomic predictions for new zealand dairy bulls and integration with national genetic evaluation. *Journal of Dairy Science*, 93(3):1243 – 1252.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, 9(2):226–252.

Henderson, C. R. (1963). Selection index and expected genetic advance. *Statisitical Genetics and Plant Breeding*.

Henderson, C. R. and Quaas, R. L. (1976). Multiple trait evaluation using relatives' records. *Journal of Animal Science*, 43(6):1188–1197.

Hill, W. (2010). Understanding and using quantitative genetic variation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365:73–85.

Jia, Y. and Jannink, J.-L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, 192(4):1513–1522.

Klemetsdal, G. (1992). Estimation of genetic trend in racehorse breeding. *Acta Agriculturae Scandinavica, Section A — Animal Science*, 42(4):226–231.

Koerhuis, A. and van der Werf, J. (1994). Uni- and bivariate breeding value estimation in a simulated horse population under sequential selection. *Livestock Production Science*, 40(2):207 – 213.

Kruuk, L. E. B. (2004). Estimating genetic parameters in natural populations using the animal model. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1446):873–890.

Liao, J. G. and Lipsitz, S. R. (2002). A type of restricted maximum likelihood estimator of variance components in generalised linear ixed models. *Biometrika*, 89(2):401–409.

Lynch, M., Walsh, B., et al. (1998). *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA.

Mehrabani-Yeganeh, H., Gibson, J., and Uimari, P. (1999). The effect of using different culling regimens on genetic response with two-trait, two-stage selection in a nucleus broiler stock. *Poultry Science*, 78(7):931 – 936.

Meuwissen, T., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.

Mrode, R. (2014). *Linear Models for the Prediction of Animal Breeding Values: 3rd Edition*. CABI.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554.

Pollak, E., Van der Werf, J., and Quaas, R. (1984). Selection bias and multiple trait evaluation. *Journal of Dairy Science*, 67(7):1590 – 1595.

Postma, E. (2006). Implications of the difference between true and predicted breeding values for the study of natural selection and micro-evolution. *Journal of Evolutionary Biology*, 19(2):309–320.

Robinson, G. K. et al. (1991). That blup is a good thing: the estimation of random effects. *Statistical science*, 6(1):15–32.

Searle, S. R. (1995). An overview of variance component estimation. *Metrika*, 42(1):215–230.

Song, H., Zhang, J., Zhang, Q., and Ding, X. (2019). Using different single-step strategies to improve the efficiency of genomic prediction on body measurement traits in pig. *Frontiers in Genetics*, 9:730.

Sorensen, D., Wang, C., Jensen, J., and Gianola, D. (1994). Bayesian analysis of genetic change due to selection using gibbs sampling. *Genetics Selection Evolution*, 26(4):333.

Stock, K., Distl, O., and Hoeschele, I. (2008). Bayesian prediction of breeding values for multivariate binary and continuous traits in simulated horse populations using threshold–linear models with gibbs sampling. *animal*, 2(1):9–18.

Van Tassell, C. P. and Van Vleck, L. D. (1996). Multiple-trait gibbs sampler for animal models: flexible programs for bayesian and likelihood-based (co) variance component inference. *Journal of Animal Science*, 74(11):2586–2597.

Veatch-Blohm, M. E. (2007). Principles of plant genetics and breeding. *Crop Science*, 47(4):1763–1763.

Volpato, L., da Silva Alves, R. A., Teodoro, P. E., de Resende, M. D. V., Nascimento, M., Nascimento, A. C. C., Ludke, W. H., da Silva, F. L., and Borém, A. (2019). Multi-trait multi-environment models in the genetic selection of segregating soybean progeny. *PLoS ONE*, 14(4):e0215315.

Weller, J. I. (2016). *Maximum Likelihood, Restricted Maximum Likelihood, and Bayesian Estimation for Mixed Models*, chapter 8, pages 51–58. John Wiley Sons, Ltd.

West, B. T., Welch, K. B., and Galecki, A. T. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman and Hall/CRC, second edition.